2020

Per-voxel Prediction on Medical Images via Deep Neural Networks

Biting Yu

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Per-voxel Prediction on Medical Images via Deep Neural Networks

Biting Yu

This thesis is presented as part of the requirements for the conferral of the degree:

Doctor of Philosophy

Supervisor:
A/Prof. Lei Wang

Co-supervisor:
Dr. Luping Zhou

The University of Wollongong
School of Computing and Information Technology

December, 2020

Certification

I, *Biting Yu*, declare that this thesis is submitted in partial fulfilment of the requirements for the conferral of the degree *Doctor of Philosophy*, from the University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Biting Yu

December 14, 2020

Abstract

Medical image synthesis and segmentation are two essential per-voxel prediction tasks, which can be applied in various clinical situations and can assist professional experts to solve various practical problems. Both of these two tasks need to estimate the target value of each voxel in input medical images. For synthesis, the target value is the intensity in the target image; for segmentation, the target value is the class label indicating whether the voxel belongs to the region of interest or not. Compared with many other computer vision tasks, like image classification that estimates image-level label, high-quality medical image synthesis and segmentation require to capture more visual details of both local and global context to predict the dense voxel information. In recent years, various deep neural network models have been developed to solve data processing problems, as they possess the powerful capacity of extracting task-specific features from input data. Among these models, deep convolutional neural networks (CNNs) have shown their promising performance in the field of computer vision. For generic image per-pixel prediction tasks, the delicately designed CNNs can capture the crucial underlying features that represent both the local and global knowledge from given images and efficiently estimate their corresponding target outputs. Owing to the astonishing learning capability of CNNs, this thesis aims to explore more effective and efficient deep CNNs based methods to solve the medical image per-voxel prediction problems. Since medical images have their own characteristics, such as higher image dimensions and less accessible labeled data than generic images, it is challenging to design CNNs based models to fully and explicitly exploit these characteristics and address the learning issues caused by them. Therefore, this thesis also focuses on exploring more advanced learning techniques and integrating them into the learning of deep CNNs to further improve the per-voxel prediction performance on medical images. Specifically, this thesis unfolds its investigation on medical image synthesis and segmentation from the following four aspects.

Firstly, this thesis develops adversarial learning based deep CNN models for cross-modality magnetic resonance (MR) image synthesis. Although deep CNNs have the dominant strength in image feature extraction, the CNNs based generative adversarial networks (GANs) have shown more promising performance for generic image synthesis recently. With the adversarial competition between the generator and discriminator, GANs can synthesize more realistic images than the conventional CNNs. However, if di-

rectly applying these GANs on medical image synthesis, the final results will not meet the expectation. One of the reasons is that many medical images, such as MR images, have three dimensions. The 2D GANs that are commonly used on generic images easily fail to capture the continuous visual clues across the 2D slices of the input MR images. To deal with this problem, 3D CNNs based GANs model is developed in this thesis to learn the synthesis mapping from one MR modality to another. Also, the designed GANs model uses the Unet-like generator so that both of the local object content and the whole image context from the given images can be seized in a larger 3D scope for better synthesis. The proposed 3D GAN model is demonstrated to be superior over the general 2D GANs on a public MR image dataset.

Secondly, after proposing the effective 3D GAN model for MR image synthesis, this thesis points out that compared with generic image synthesis, medical image synthesis needs more efforts to depict the textural structures of interesting objects, as this structural information is crucial for accurate disease diagnosis and other recognition tasks. Whereas, the above 3D GAN model and most existing GANs only attempt to minimize the pixel-/voxel-wise intensity distance between the real and the synthesized images during training, which is insufficient to preserve the vital structural details. Since image edge information can reflect the textural structure of image content and depict the boundaries of different objects in images, this thesis further explores two learning strategies to integrate the edge information into the adversarial learning of GANs. Using the adversarially learnt edge maps, the proposed models could enforce both voxel-wise intensity similarity and edge similarity between the real and synthesized images and ensure the sharpness of the predicted images. The superiority of the proposed 3D edge-aware GANs is demonstrated on various public MR brain image datasets.

Thirdly, although the above 3D edge-aware GANs can effectively predict sharper images, they just learn a single model to uniformly transform all the input images by a whole sample-space mapping, as most existing CNNs based methods do. In this thesis, it is argued that this may not be sufficient for medical image synthesis as the limited labeled training data are often not representative enough to cover the variations in the unseen images. Thus, it is difficult in utilizing these labeled images to train a single optimal mapping model for all the images. To handle this issue, this thesis develops a novel GANs based sample-adaptive learning framework. It seeks both of the common whole sample-space mapping between the source- and target-modalities and an additional unique local sample-space mapping for each input sample via exploring its specific characteristic. Specifically, the learning model is decoupled into two intercommunicated paths. In its baseline path, the global sample-space mapping is learnt as usual to fit all the available labeled samples by a common GAN model. At the same time, a new sample-adaptive path is designed to further learn the relationship between each input sample and its neighboring training samples and exploits the target-modality features of these training samples as auxiliary

information for synthesis. Benefiting from this sample-adaptive learning strategy, the proposed GANs based model possesses the flexibility to adapt itself to different samples so that the synthesis performance can be improved. The effectiveness of the proposed sample-adaptive learning framework is validated with two different GANs on two lesion contained MR image datasets.

Lastly, this thesis further explores sample-adaptive learning for brain tumor segmentation. Since MR images are not quantitative during imaging and can exhibit significant variations in signal depending on a range of factors, it experiences an increasing difficulty to train an automatic segmentation network and apply this trained network to new MR images. To mitigate this issue, this thesis proposes to learn a sample-adaptive intensity lookup table (LuT) that dynamically transforms the intensity contrast of each input MR image to adapt to the subsequent segmentation task. To be specific, the proposed sample-adaptive framework contains a LuT module and a segmentation module, trained in an end-to-end manner: the LuT module learns a sample-specific nonlinear intensity mapping function through communication with the segmentation module, targeting to improve the ultimate segmentation performance. In order to make the LuT module sample-adaptive, the intensity mapping function is parameterized by exploring two families of non-linear functions. The parameters of these functions are specifically predicted for each input sample, making the intensity mapping adaptive to samples. With this sample-adaptive learning, the final segmentation performance can be boosted. The proposed framework is developed upon two state-of-the-art backbone networks for segmentation. Its effectiveness is validated on two benchmark brain tumor segmentation datasets. The experimental results indicate that rather than learning the segmentation-model specific information, the LuTs learnt in our approach also carry the general information about the intensity level adjustment for the given segmentation task.

Acknowledgments

My PhD study has lasted for about four years. It is not just an individual experience. During this period, I have received so much kind help from many people. In this thesis, I would like to thank them sincerely.

First of all, I want to express my deepest appreciation to my supervisors, Dr. Luping Zhou and Dr. Lei Wang. Because of them, I realize that being a researcher requires passion, hardworking, carefulness, and the most important thing, devotion. They spent a lot of time guiding me to explore my research topic step by step. They shared new ideas with me, helped me to solve practical problems, and taught me how to present the works. I will never forget each time before the paper submission deadline, sometimes they were sick, they would still focus on revising and polishing my papers again and again. Without their support, it is not possible for me to finish any single work in these four years.

I am also extremely grateful to my joint supervisors from CSIRO, Dr. Pierrick Bourgeat and Dr. Jurgen Fripp. They have provided me many valuable suggestions in my study. The suggestions, especially those relating to the clinical practice, have largely improved my research works. Special thanks should then go to the other co-authors of my published works, Dr. Yinghuan Shi, Dr. Wanqi Yang, and Dr. Ming Yang. Their help further enhanced the quality of these works.

I would like to extend my sincere thanks to my friends and lab-mates in the University of Wollongong, Dr. Peng Wang, Dr. Zhimin Gao, Dr. Yan Zhao, Dr. Yan Wang, Dr. Huan Wang, Dr. Lei Qi, Dr. Yang Li, Dr. Jiashuang Huang, Dr. Chen Zu, Zhongyan Zhang, Yu Ding, Saimunur Rahman, Din Sangrasi, Zhexuan Zhou, Melih Engin, Ian Comor, Bela Chakraborty, Xishun Wang, Huaming Chen, Jianqing Wu, Xiwu Zhang and so on. Thank you very much for your help and company.

Many thanks to Mr. Jun Hu and Ms. Yuan Tian. They helped me to solve computer and server problems. Without their IT support, I cannot complete the experiments on remote servers.

My sincere thanks go to my dear family. My parents and parents-in-law never waver in their support to my life and study. The physical distance between us never stopped them from expressing their love to me. Whenever I was sad or exhausted, they were waiting at the other end of internet to comfort and encourage me. My husband, Bo Jia, puts my personal feelings in the first place. His love supports me to overcome all the difficulties.

When I felt tired of study, he always told me to take care of my physical and emotional health instead of continuing working. He often said that my happiness was much more important than the PhD degree. As he never gave me any pressure, I could enjoy my daily life after working.

I must thank myself. My optimistic and positive attitude encouraged me to go through the hardships of study. My determination backed me up in every working night. My cooking skills healed the homesickness in these four years. I also thank this oversea study experience. It has prepared me as a stronger and more independent person for the following life.

Finally, I would like to thank University of Wollongong for the offered international postgraduate tuition award and DECRA scholarship and CSIRO for the top-up living allowance to support my study in Australia.

Publications

This thesis consists of the content in the consistent order with the following first-authored publications/manuscripts:

- **Biting Yu**, Yan Wang, Lei Wang, Dinggang Shen, and Luping Zhou. "Medical Image Synthesis via Deep Learning." In Deep Learning in Medical Image Analysis, pp. 23-44. Springer, Cham, 2020.
- **Biting Yu**, Luping Zhou, Lei Wang, Jurgen Fripp, and Pierrick Bourgeat. "3D cGAN based cross-modality MR image synthesis for brain tumor segmentation." In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 626-630. IEEE, 2018.
- **Biting Yu**, Luping Zhou, Lei Wang, Yinghuan Shi, Jurgen Fripp, and Pierrick Bourgeat. "Ea-GANs: edge-aware generative adversarial networks for cross-modality MR image synthesis." IEEE transactions on medical imaging (IEEE TMI) 38, no. 7 (2019): 1750-1762.
- **Biting Yu**, Luping Zhou, Lei Wang, Yinghuan Shi, Jurgen Fripp, and Pierrick Bourgeat. "Sample-adaptive GANs: Linking Global and Local Mappings for Cross-modality MR Image Synthesis." IEEE transactions on medical imaging (IEEE TMI) 39, no. 7 (2020): 2339-2350.
- **Biting Yu**, Luping Zhou, Lei Wang, Wanqi Yang, Ming Yang, Jurgen Fripp, and Pierrick Bourgeat. "Learning Sample-adaptive Intensity Lookup Table for Brain Tumor Segmentation." Accepted by 2020 23th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI2020).
- **Biting Yu**, Luping Zhou, Lei Wang, Wanqi Yang, Ming Yang, Jurgen Fripp, and Pierrick Bourgeat. "SA-LuT-Nets: Learning Sample-adaptive Intensity Lookup Tables for Brain Tumor Segmentation." Submitted to IEEE transactions on medical imaging (IEEE TMI).

Except for the first-authored publications listed above, the following co-authored publications are also developed during my Ph.D. course. The main body of this thesis does not include them.

- Yan Wang, **Biting Yu**, Lei Wang, Chen Zu, David S. Lalush, Weili Lin, Xi Wu, Jiliu Zhou, Dinggang Shen, and Luping Zhou. "3D conditional generative adversarial networks for high-quality PET image estimation at low dose." Neuroimage 174 (2018): 550-562.

- Yan Wang, **Biting Yu**, Lei Wang, Chen Zu, Yong Luo, Xi Wu, Zhipeng Yang, Jiliu Zhou, and Luping Zhou. "Tumor segmentation via multi-modality joint dictionary learning." In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1336-1339. IEEE, 2018.

- Yan Wang, Luping Zhou, **Biting Yu**, Lei Wang, Chen Zu, David S. Lalush, Weili Lin, Xi Wu, Jiliu Zhou, and Dinggang Shen. "3D auto-context-based locality adaptive multi-modality GANs for PET synthesis." IEEE transactions on medical imaging 38, no. 6 (2018): 1328-1339.

- Moghari, M. Dashtbani, Luping Zhou, **Biting Yu**, K. Moore, N. Young, R. Fulton, and A. Kyme. "Estimation of full-dose 4D CT perfusion images from low-dose images using conditional generative adversarial networks." In 2019 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), pp. 1-3. IEEE, 2019.

- Yan Wang, Luping Zhou, Lei Wang, **Biting Yu**, Chen Zu, David S. Lalush, Weili Lin, Xi Wu, Jiliu Zhou, and Dinggang Shen. "Locality adaptive multi-modality GANs for high-quality PET image synthesis." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 329-337. Springer, Cham, 2018.

List of Abbreviations

- CT: computerized tomography
- PET: positron emission tomography
- MRI: magnetic resonance imaging
- CNNs: convolutional neural networks
- GANs: generative adversarial networks
- cGANs: conditional GANs
- WGAN-GP: Wasserstein GAN with gradient penalty
- Conv: convolutional
- FC: fully connected
- ReLUs: rectified linear units
- SGD: stochastic gradient descent
- ResNet: residual neural network
- FCN: fully convolutional network
- 2D: two-dimensional
- 3D: three-dimensional
- CRFs: conditional random fields
- Ea-GANs: edge-aware GANs
- gEa-GAN: generator-induced Ea-GAN
- dEa-GAN: discriminator-induced Ea-GAN
- SA-GANs: sample-adaptive GANs

- LuTs: lookup tables
- SA-LuT-Net: sample-adaptive lookup table network
- PSNR: peak signal-to-noise ratio
- NMSE: normalized mean squared error
- SSIM: structural similarity index
- STD: standard deviation

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Background

Medical imaging techniques, such as computerized tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI), play a crucial role in clinics. Doctors usually employ them to visualize the internal structures or functions of human bodies via medical images. These images varying with imaging techniques and scanner settings can highlight diverse anatomical organs and physiological tissues. Hence, to better serve the disease diagnosis and treatment, multiple modalities of medical images are collected together to reflect the different and sometimes complementary visual specialty of body parts of interest. However, due to the uneven clinical resources and expensive acquisition costs, the demand of multiple and also high-quality medical images is not easily satisfied for every patient in the practice. This always results in the less comprehensive knowledge about diseases and causes inaccurate diagnosis. Medical image synthesis, as an approach modelling a mapping from a given modality of images to a target modality, is therefore widely explored to handle this issue. Meanwhile, to assist doctors in diagnosis and treatment, the accurate appearance and location information about organs and abnormal tissues is important. This information is commonly acquired through segmenting these target objects on the scanned medical images. Whereas, different from the segmentation labelling on generic images, the manual annotation on medical images requires laborious efforts from the medical experts to delineate the contours of target objects. Therefore, to release the workload in clinics, the more effective and efficient approaches are always in high demand for automatic medical image segmentation.

Both medical image synthesis and segmentation belong to an essential computer vision technique that is known as per-pixel (per-voxel for 3D medical images) prediction, as shown in Figure 1.1 (1). Medical image synthesis corresponds to per-voxel regression as it estimates the intensity value of each voxel in the target-modality images, while medical image segmentation corresponds to per-voxel classification as it categories every voxel

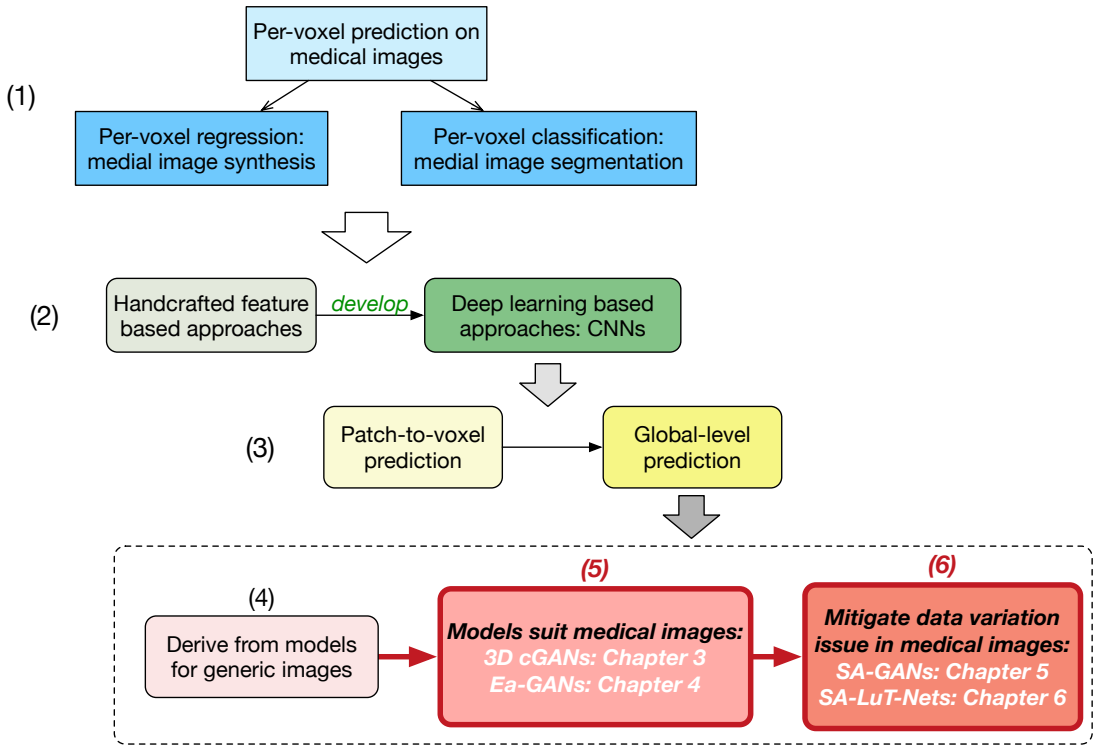


Figure 1.1: Overview: (1) per-voxel prediction methods on medical images can be classified into per-voxel regression (image synthesis) and per-voxel classification (image segmentation); (2) methods for per-voxel prediction develop from using handcrafted features to learning task-specific features by CNN models; (3) CNNs based models are explored from patch-to-voxel prediction to global-level prediction; (4) most existing per-voxel prediction methods on medical images directly derive from the models used on generic images; (4) this thesis will design effective deep CNNs which can extract volumetric and object structure related information from 3D medical images; (4) this thesis will develop sample-adaptive deep CNNs to mitigate the data variation issue in learning the sample-unified models.

into the object of interest or the background. Therefore, in these two challenging tasks, the predicted results have the same spatial size with the input images. Per-voxel prediction is a more difficult undertaking than image classification that only needs to estimate a single value (label) for each entire image. Successful per-voxel prediction systems highly rely on their understanding about the whole given images and require the effective feature representation about object locations, appearance, and image context. Conventionally, handcrafted image features are delicately selected for different tasks, then the selected features are processed by machine learning models for per-voxel prediction. In the recent decade, thanks to the improvement of hardware, deep neural networks based models can learn the optimal features from images at the same time with the prediction, bringing out a breakthrough in the computer vision area [1], as shown in Figure 1.1 (2). Among these models, deep convolutional neural networks (CNNs) demonstrate their domination with the superior performance in various generic image recognition tasks, such as object detection [2], semantic segmentation [3], and face recognition [4]. The tasks on medical images also have witnessed the successful applications of CNNs with their remarkable capability in feature extraction from multi-dimensional data [5].

In the beginning, CNNs based methods were explored for per-voxel prediction tasks in medical image analysis using the patch-to-voxel prediction strategy. For example, a CNN model, stacked by convolutional layers and fully connected layers, processes each local patch from MR images to label its centered voxel one-by-one via sliding windows for brain tumor segmentation in [6]. Since this kind of patch-to-voxel prediction strategy ignores the important context information among different patches in the same image, an independent conditional random field (CRF) is often added after the CNNs to model the relationship among all the pixels in an image. Whereas, this patch-to-voxel prediction has low efficiency, and the additional CRF learnt as a separate step cannot take the advantages from the whole deep learning process. To cope with these issues, fully convolutional networks (FCNs) were developed and made the global-level prediction possible in generic semantic image segmentation [7], as shown in Figure 1.1 (3). Owing to their successful applications, diverse FCN architectures have been proposed to directly predict their corresponding target images or segmentation labels from the given whole images in per-voxel prediction tasks. Using more advanced model structures, like Unet [8], FCN can not only seize the image information about local objects and their global relationship at the same time, but also process every input image via the one-pass inference to improve the per-voxel prediction efficiency. Thus, the deep FCNs, which can grasp both local and global image representations with low computational costs, are an appealing option for per-voxel prediction tasks.

Furthermore, after generative adversarial networks (GANs) achieved the successful performance in generic image synthesis [9], the CNNs trained by adversarial learning started to be explored in per-voxel prediction on medical images [10]. Different from the

conventional CNNs only focusing on prediction, a CNN based GAN model benefits from the learning competition between its generator and an additional discriminator. During this competition, the per-voxel prediction output from the generator is enforced to contain sufficient content and context information to increase the discrimination difficulty of its discriminator, which in return further improves the generator to predict more realistic target-modality images or segmentation labels. However, most GAN based methods on medical images derive from the GANs that are originally proposed for 2D generic image tasks, as shown in Figure 1.1 (4). They cannot fully exploit the essential 3D nature of volumetric objects in 3D medical images, like those in CT, PET, and MR images. In addition, the crucial textural structure information about medical objects in images is not sufficiently employed by these GANs during the prediction, which lowers the quality of prediction results, such as synthesizing less sharp medical images in the per-voxel regression tasks. Therefore, it is appealing to develop more advanced and effective CNNs cooperated with the adversarial learning so that they can better suit the processing on medical images, as shown in Figure 1.1 (5).

Besides, despite the various architectures used in per-voxel prediction methods, the existing CNNs usually attempt to train a unified model that processes all unseen images. Successfully training this unified CNN model requires a large number of labeled images. However, sufficient labeled medical images are often inaccessible, since labeling these images or scanning new images relies on the experts with professional background and adequate clinical resources. Such a situation is even more protruding in per-voxel prediction tasks where the value of every voxel needs to be predicted. The labeled training images are in a small number and less representative to a varied population, making it hard to train a unified CNN model to fit all test images. This issue will become more severe when the visual variation is significant among medical images. For example, since the intensity values in MR images are not quantitative, the scanned intensities of the same tissue type can be distinctly different among MR images. If the MR images contain tumor lesions, like gliomas, the tumor-surround-contrast can be much different from some MR images to the others. Also, the tentacle-like structures of gliomas often stretch to invade the normal brain tissues, rather than just replacing them, which leads to the arbitrary appearance and diffused locations of tumors on the scanned images varying from patient to patient [11]. The arbitrary variations among these medical images further increase the difficulty to learn an optimal unified model for all samples for per-voxel prediction. A few works attempted to address this variation problem using the domain adaptation strategy. They usually tried to reduce the variation between the whole training set and the whole test set so that the trained model can adapt to the test set [12, 13]. These methods assumed that the training and test data were from two different distributions and minimized the discrepancy between them. Whereas, these domain adaptation approaches ignore the different characteristics among the data in the same set. Therefore, this strategy from the

literature has not directly addressed the data variation problem among all the images in learning a unified per-voxel prediction model. It is appealing to explore the CNN based methods that can adapt to the different image samples and solve this research problem in a straight-forward way for the per-voxel prediction tasks, as shown in Figure 1.1 (6).

1.2 Research Aims

This thesis focuses on developing effective deep neural networks, especially the powerful CNNs based models, for per-voxel prediction on medical images. The developed CNNs models target at incorporating 1) adversarial learning that can capture the volumetric local and global features from medical images and fully exploit their object structure information during the prediction and 2) sample-adaptive learning that can mitigate the data variation issue in learning the sample-unified models, so that the performance of these CNN models can be well generalized. To be more specific, this thesis has the following three research aims.

- Design effective deep CNNs based GAN models to learn the medical image mapping for per-voxel regression tasks, e.g., cross-modality MR image synthesis. The designed model should be able to extract the volumetric details of objects and also the contextual information in the whole images. During the adversarial learning of the designed model, it should exploit the sufficient structural details of objects and preserve it in the synthesized images to enhance their sharpness. Besides, the designed model should get better performance on various datasets than the state-of-the-art approaches using either handcraft features, conventional deep CNNs, or GANs.

- Develop a GANs based sample-adaptive learning framework that can handle the aforementioned data variation issue in learning a unified prediction model for all samples in the complicated per-voxel regression tasks, e.g., cross-modality lesion-contained MR image synthesis. The developed sample-adaptive framework should be able to extract the characteristic of each individual sample and adjust the learning models based on different input samples. This learning framework needs also to be flexibly developed on different GANs and cooperate the sample-adaptive learning with them to improve their corresponding synthesis performance. Besides, this GANs based framework should better perform on various lesion contained datasets compared with other advanced GAN models without the sample-adaptive learning.

- Develop a novel deep CNNs based learning framework that can vary with samples to mitigate the significant visual variation among different medical images for the difficult per-voxel classification tasks, e.g., brain tumor segmentation on MR

images. This framework should learn to build the sample-adaptive visual transformation for each input image and adapt this learnt specific transformation to the segmentation task to improve its performance. It also should be able to work upon different segmentation networks and outperforms the state-of-the-art deep CNNs without the sample-adaptive learning on benchmark datasets.

1.3 Contributions

To dedicate the research to the above three aims, the contributions of this thesis are highlighted as follows.

- This thesis proposes a 3D CNNs based GAN model, i.e., 3D cGAN, to learn the mapping for cross-modality brain MR image synthesis at the global level. With the 3D Unet-like structure in its generator, it captures the volumetric features of both local object content and whole image context in a larger 3D scope for synthesis. Furthermore, a local adaptive fusion approach is additionally proposed to polish the synthesized MR images from 3D cGAN. This thesis demonstrates the better MR image synthesis performance of the proposed 3D cGAN than that of the general 2D cGAN model on a public brain tumor MRI dataset, i.e. BRATS2015 [14], and also validates the effectiveness of the local adaptive fusion strategy to polish the final synthesis.

- This thesis points out that purely enforcing the voxel-wise intensity similarity is not sufficient for medical image synthesis. Thus, it further develops edge-aware generative adversarial networks (Ea-GANs) for cross-modality brain MR image synthesis. The proposed Ea-GANs are designed to preserve the edge information that can reflect the brain structure as the object contour information in images to improve the sharpness of synthesized images. To integrate the edge information, two variants of Ea-GANs, i.e., generator-induced Ea-GAN (gEa-GAN) and discriminator-induced Ea-GAN (dEa-GAN), are proposed according to different learning strategies. In the gEa-GANs, edge information is incorporated into the adversarial learning of the generator, enforcing the synthesized image to have a similar edge map as the real image. In the dEa-GAN, the edge information is innovatively incorporated into both of its generator and discriminator. In this way, the edge information will also be adversarially learnt, which could further preserve the object-related information in the synthesized image to improve the synthesis performance. The proposed Ea-GANs are validated on two public MRI datasets, i.e. the brain tumor contained BRATS2015 [14] and the non-skull-stripped IXI [15], respectively. The experimental results demonstrate the superior performance of the proposed Ea-GANs

over a set of state-of-the-art image synthesis models including the model using handcrafted features, conventional deep CNNs, and other GANs models.

- This thesis proposes a novel GANs based sample-adaptive learning framework, called SA-GANs, to enforce the sample-specific learning on top of the common whole sample-space learning, so that the unique characteristic of each medical image can be also seen and exploited by the proposed framework for the cross-modality MR image synthesis. To be more specific, the proposed framework decomposes the learning process into two cooperated paths. In the baseline path, the global sample-space mapping is learnt from the whole source-modality space to the target-modality space by a GAN model as usual. Additionally, it has a sample-adaptive path to seek the characteristic of each individual sample through learning its unique local sample-space mapping. This sample-specific path models the relationship between each given sample and its neighboring training samples and utilizes the target-modality features of these training samples as auxiliary information to promote the final high-quality synthesis. The proposed sample-adaptive learning framework can be separately developed on two different GANs and end-to-end trained with them. Its effectiveness is validated on two MR image datasets, i.e., brain tumor contained BRATS2015 [14] and stroke lesion contained SISS2015 [16]. The experimental results show the better performance of the proposed SA-GANs than a set of state-of-the-art GAN models that are trained without sample-specific learning.

- This thesis proposes a CNNs based sample-adaptive learning framework called SA-LuT-Nets for segmentation where intensity lookup tables (LuTs) are learnt in an end-to-end manner with a subsequent segmentation network to cope with the significant variation among MR images and promote brain tumor segmentation performance. The learnt LuTs vary with the need of the input MR images during the intensity transformation for better segmentation. In this way, when the new unseen samples whose intensity distributions are different from the training set arrive, our SA-LuT-Nets could predict their specific optimal LuTs to adjust their intensities for the segmentation. The proposed framework is developed and validated based on two segmentation backbones, i.e., the modified 3D Unet [17] and DMFNet [18], which achieved the state-of-the-art performance on brain tumor segmentation. The effectiveness of our proposed learning framework over these two baselines is demonstrated in both single- and multi-modalities scenarios on the two public brain tumor segmentation datasets, i.e., BRATS2018 and BRATS2019 [19]. The online evaluation results validate that our SA-LuT-Nets achieves better performance than many other state-of-the-art CNNs models, while using fewer model parameters. The thesis also shows that, the LuTs learnt using one segmentation

model could be transferred to another segmentation model to improve the latter's performance. This suggests that some general information about how to sample-adaptively adjust the intensity levels of MR images for the segmentation task, rather than the information merely coped with a specific segmentation model, has been learnt.

Briefly, this thesis has conducted research into establishing effective deep CNNs based models for the two challenging per-voxel prediction tasks, i.e., medical image synthesis and segmentation, through the adversarial learning and sample-adaptive learning techniques. The research starts from the aim of designing effective CNNs based GAN models for per-voxel synthesis on 3D MR images. The investigation goes through developing the CNN architectures to learn the volumetric feature representations of medical images. After mitigating the discontinuous cross-slice estimation by the 3D-structure model, this thesis further integrates the edge information into the adversarial learning to preserve the essential structural texture of brains for the sharper images after the synthesis. After solving the first research aim by the more effective CNNs based models for per-voxel prediction, the research comes into dealing with the data variation problem in learning a unified model for medical images. It explores sample-adaptive learning based frameworks to handle this research problem for both synthesis and segmentation tasks. A sample-adaptive learning framework is developed to learn a sample-specific mapping in addition to a unified mapping to capture the characteristic of each sample during synthesis. Besides, sample-adaptive learning is explored to dynamically adjust the intensity contrast to handle the significant visual variation among MR images before processing them by a segmentation network for the ultimate task. These two studies have separately realized the second and the third research aims in this investigation. Different experiments have been conducted, and their results have validated the effectiveness of the proposed special design in the deep CNN based models, including the 3D cGAN, Ea-GANs, and SA-GANs for cross-modality MR image synthesis, and the SA-LuT-Nets for brain tumor segmentation.

1.4 Organization of the Thesis

The following part of this thesis will be organized as follows.

Chapter 2 first reviews the classic architectures of deep CNNs models and the learning approaches of GAN models for per-voxel prediction tasks. Then, the existing key works about CNNs based methods targeting at the per-voxel regression and classification tasks, i.e. medical image synthesis and segmentation, are separately reviewed.

Chapter 3 points out the weakness of the existing 2D CNNs based GAN models in synthesizing 3D medical images and introduces the proposed 3D cGAN model for brain MR

image synthesis. The superiority of its 3D architecture over the 2D structure is demonstrated on a brain tumor MR image dataset.

Chapter 4 presents the proposed Ea-GANs models for the cross-modality MR image synthesis task. Its design motivation and model details are investigated and provided. The effectiveness of the proposed two edge-aware learning strategies is validated on two datasets compared with the 2D CNNs and GAN models and the state-of-the-art MR image synthesis methods.

Chapter 5 introduces the importance of building the sample-specific learning for the successful lesion contained MR image synthesis GAN models. The proposed sample-adaptive learning framework that learns the specific local sample-space mapping for each sample are described, and its effectiveness is verified on two lesion contained MR image datasets compared with the other advanced and popular GANs without using this learning strategy.

Chapter 6 reveals the intensity variation issue of MR images in the existing automatic brain tumor segmentation methods. The approach of employing intensity lookup tables to handle this issue is provided. The details about using the proposed framework to conduct the sample-adaptive intensity adjustment for the better segmentation are then introduced. The experimental studies on both single-modality and multi-modality cases are presented to demonstrate its improvement over the two different CNN segmentation baselines, as well as other prevalent brain tumor segmentation models.

Chapter 7 summarizes this thesis and discusses its related future works.

Chapter 2

Literature Review

In this chapter, the relevant key works about per-voxel prediction, i.e., medical image synthesis and segmentation, will be reviewed. To be more specific, the typical architectures of deep convolutional neural networks (CNNs) and the learning approaches of different generative adversarial networks (GANs) are first introduced. Then, the important medical image synthesis and segmentation methods that are based on deep CNNs will be further discussed.

2.1 Deep Convolutional Neural Networks (CNNs)

Deep learning models integrate the important feature learning into target tasks and have demonstrated themselves in many applications [20]. Among them, deep convolutional neural networks (CNNs) are frequently used especially for the computer vision tasks. A typical CNN is composed by an input layer, hidden layers, and an output layer [2]. The input layer is always associated with images, and the output layer produces their corresponding estimations. The stacked hidden layers are associated with different operations performing on the input data. According to the stacking order and operation types of these hidden layers, they can be sorted into small units in the CNN model to extract the task-related features from shallow to deep. With the increasing number of stacked units, the CNN model enhances its ability of revealing the more implicit feature representations from data. Generally, each unit at least consists of a convolutional layer and an activation layer, which can be written as the following operations:

$$\mathbf{h}_j^l = \sigma\Big(\sum_{i=1}^{N^{l-1}} \mathbf{h}_i^{l-1} * \mathbf{W}_{ij}^l + b_j^l\Big), j = 1, \ldots, N^l, \tag{2.1}$$

where $*$ and $\sigma(\cdot)$ denote a convolution operator and a nonlinear transformation, respectively. The symbols $\mathbf{W}$, b, and $\mathbf{h}$ separately indicate a convolutional kernel, an added bias, and a feature map. The number index of convolutional units in a CNN model are

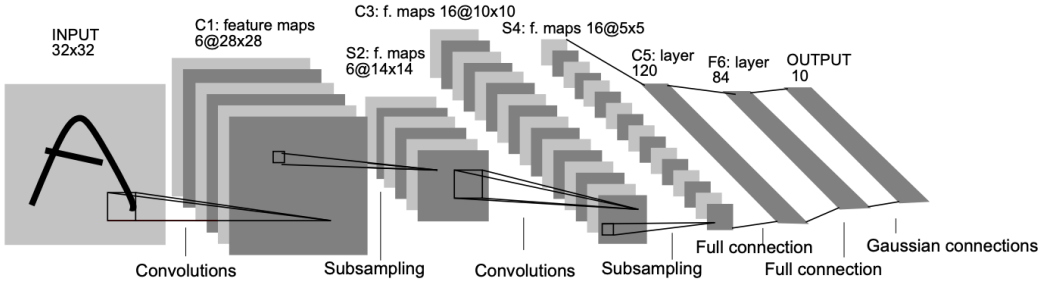


Figure 2.1: The architecture of LeNet includes three convolutional layers and two fully connected layer. Image courtesy to [21]

$1, \ldots, l, \ldots, L$ from the shallow to deep layers. N^l is the total number of feature maps in the l-th unit. Hence, Equation (2.1) presents the operating process from the feature maps generated in the $(l-1)$-th unit to the output features in the l-th unit. Except from the aforementioned convolution and activation layers, normalization, pooling, and other hidden layers are often cooperated to build a well-performed CNN. During the training of a CNN, after an input image is processed by these operations, its corresponding prediction is given by the output layer. A loss, which is defined by the vision task, calculates the error between the prediction and the real target, and then its error gradient is propagated back to the hidden layers to update the parameters in the kernels and biases for a better estimation. Using different hidden layer combinations and connection approaches, various architectures are proposed to promote the improvement of computer vision task performance. The next part will introduce eight typical CNNs architectures.

2.1.1 LeNet-5

LeNet-5, as a very early CNN model, was proposed in 1998 for handwritten character recognition task [21]. It consists of three convolutional (Conv) layers with 5 kernel size and two fully connected (FC) layers. Average pooling and tanh activation layers are also used in LeNet-5. With these layers, it processes 32×32 grayscale images of checks to recognize the digits on them. Even though its discrimination ability is limited by the shallow structure and the small number of trainable parameters, it is still a pioneering model in CNNs.

2.1.2 AlexNet

AlexNet started to attract the attention from researchers after it won ImageNet LSVRC-2012 competition with 15.3% error rate in object classification. It largely outperformed the second-place method which got 26.2%. In the AlexNet as illustrated in Figure 2.2, there are five Conv layers and three FC layers [2]. The kernel sizes are 11×11 and 5×5 correspondingly in its first and second Conv layers, and 3×3 in its last three Conv

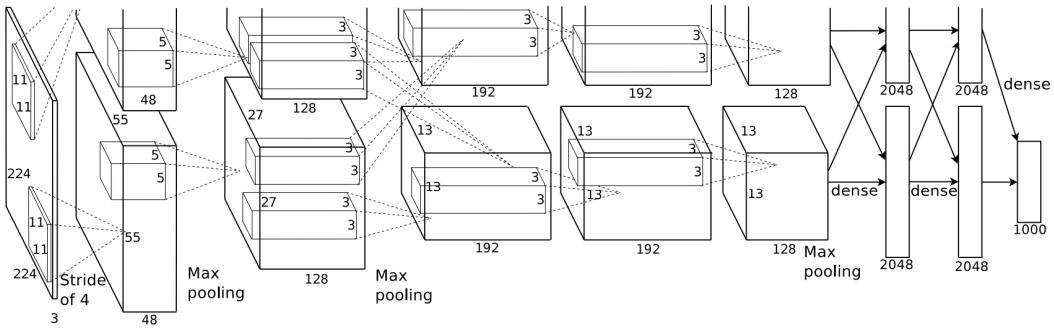


Figure 2.2: The basic architecture of AlexNet includes five convolutional layers and three fully connected layer. Image courtesy to [2]

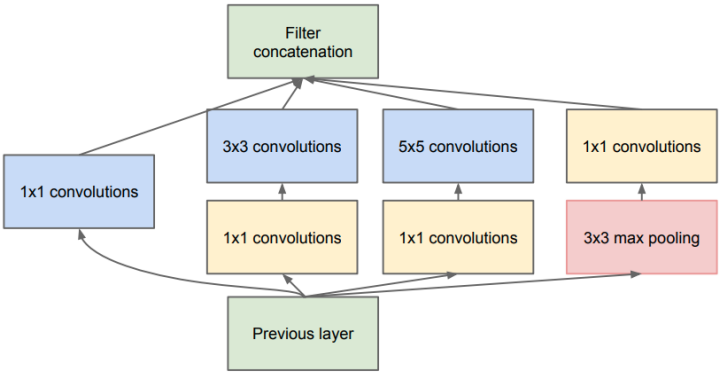


Figure 2.3: The architecture of Inception module used in GoogLeNet. Image courtesy to [23]

layers. Apart from them, it applies rectified linear units (ReLUs) [22], $f(x) = \max(0, x)$, to the output of each Conv and FC layer as nonlinear activation. The ReLU layers assist AlexNet to reach 25% training error rate by only the 1/6 training time of tanh nonlinear layers. Besides, overlapping max-pooling helps AlexNet to decrease top-1 and top-5 error rates by 0.4% and 0.3%, respectively, from the non-overlapping max-pooling operation. It also employs data augmentation on input images and random dropout on parameters during training to handle its potential overfitting. The entire model is trained by stochastic gradient descent (SGD) with momentum. All these design and discussion inspire the following development of CNNs a lot.

2.1.3 GoogLeNet

In 2014, a 22-layer deep CNN, GoogLeNet, was proposed for object classification and detection [23]. With a novel Inception CNN unit as illustrated in Figure 2.3, it reduced the top-5 error rate to 6.67% in ILSVRC-2014 as the first-place winner. The proposed Inception unit is built by four branches to get different receptive fields on input images. The receptive field of a CNN model means its seen-able spatial region in the input images and decides its ability of feature extraction. Using a small convolutional kernel size,

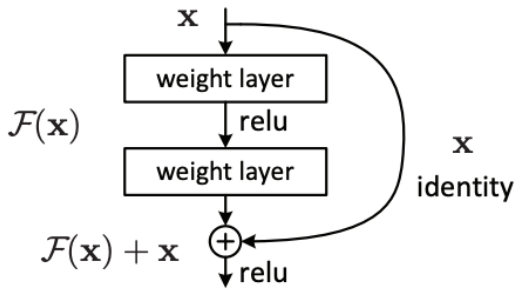


Figure 2.4: The building block of residual learning. Image courtesy to [24]

like 1×1, the CNN branch can seize the local information from images, while a larger convolutional kernel size, like 5×5, and max-pooling can help the branch to capture more global features from images. Thus, after concatenating the outputs of four branches together, Inception module can benefit from extracting both local and global image feature representations and increase the learning effectiveness. Furthermore, the designed Inception module exploits 1×1 Conv layers before 3×3, and 5×5 Conv layers to reduce the channels of feature maps so that the number of trainable parameters can be largely decreased to promote the computational efficiency. The total parameter number of GoogLeNet is about four million which is remarkably less than 60 millions of AlexNet. In addition, batch normalization, dropout, data augmentation approaches like image distortions, and RMSprop optimizer are used in the learning of GoogLeNet and make it as an effective and efficient CNN model.

2.1.4 ResNet

In 2015, residual neural network (ResNet) was introduced as the winner of ILSVRC-2015 and reached the top-5 error rate of 3.6% for classification task [24]. It incorporates a new building block called residual learning block as shown in Figure 2.4. This block is realized by a shortcut connection, which can skip one or more layers to connect to the deeper features. In the residual learning block, the shortcut connection delivers the previous feature maps by identity mapping to be added to the output feature maps from the two or three stacked layers. This process can be written as $\mathbf{y} = \mathscr{F}(\mathbf{x}, \{W_i\}) + \mathbf{x}$, where $\mathbf{x}$ and $\mathbf{y}$ indicate the input and final output features of the residual learning block, respectively, and $\mathscr{F}(\mathbf{x}, \{W_i\})$ means the learnt residual mapping with parameters $\{W_i\}$. Stacking this residual learning blocks, ResNet can be built to 152 layers almost without degradation. Degradation is a common issue, appearing when a very deep CNN begins to converge during training. If add more layers to deep CNNs, some of them cannot improve their prediction accuracy or even degrade their performance quickly since their accuracy is already saturated. In ResNet, using the residual learning blocks effectively alleviates the degradation issue, which makes ResNet as another milestone in the development of

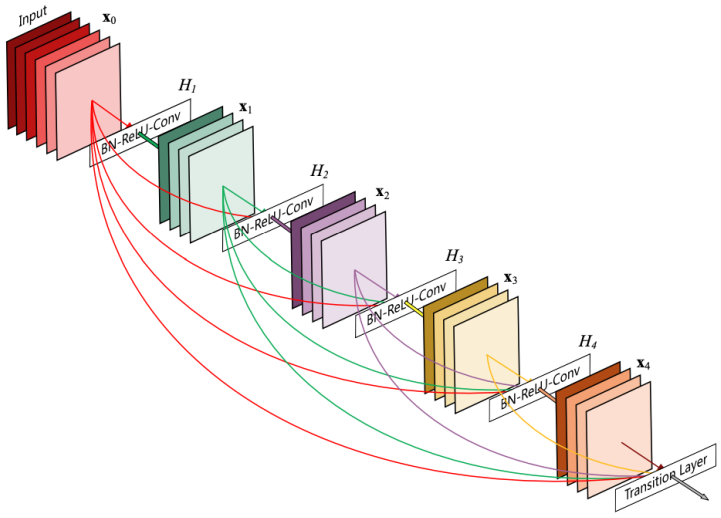


Figure 2.5: The architecture of 5-layer dense block in DenseNet. Image courtesy to [25]

CNNs.

2.1.5 DenseNet

DenseNet is also a prevalent CNN model for object recognition [25]. It consists multiple novel dense blocks as shown in Figure 2.5. As mentioned, in ResNet, identify mapping is presented to help the backpropagation of error gradients. Its shortcut connection is applied as the element-wise summing between the previous features and the features after stacked Conv layers. Differently, in the dense block, skip connection is employed as the feature concatenation. Besides, the feature maps from one layer will be transited to all its subsequent layers in the same block. Hence, after stacking multiple dense blocks with the interval of 1×1 Conv layers and 2×2 average-pooling layers, the error gradients can be easily back-propagated to every layer in the entire DenseNet, which is a compact model rather than a wide one. It achieved impressive performance in classification competitions with less trainable parameters. However, since the feature maps of all the layers in dense blocks need to store, quadratic memory should be used during the training of DenseNet.

2.1.6 Fully Convolutional Networks (FCN)

Fully convolutional networks (FCN) was elaborated to solve the pixel-wise classification, i.e., segmentation problem [7]. It converts the CNN classifiers, like AlexNet and GoogLeNet, to segmentors through replacing all the fully connected layers by the convolutional layers. In this way, FCN can keep the multi-dimensional content information with less trainable parameters and inference time. Its detailed architecture is illustrated in Figure 2.6. As can be seen, downsampling operations are first applied to abstract the semantic features, and upsampling is used to reconstruct the local content information.

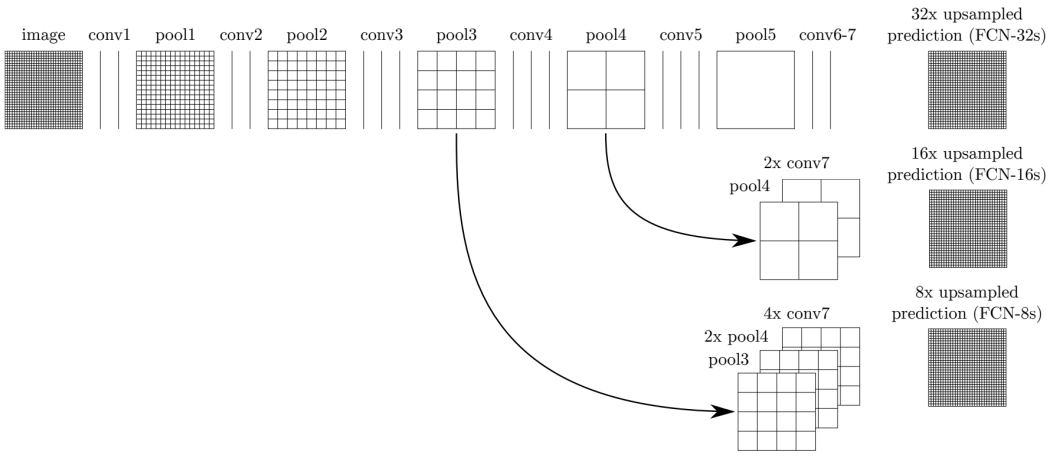


Figure 2.6: The basic architecture of FCN. Image courtesy to [7]

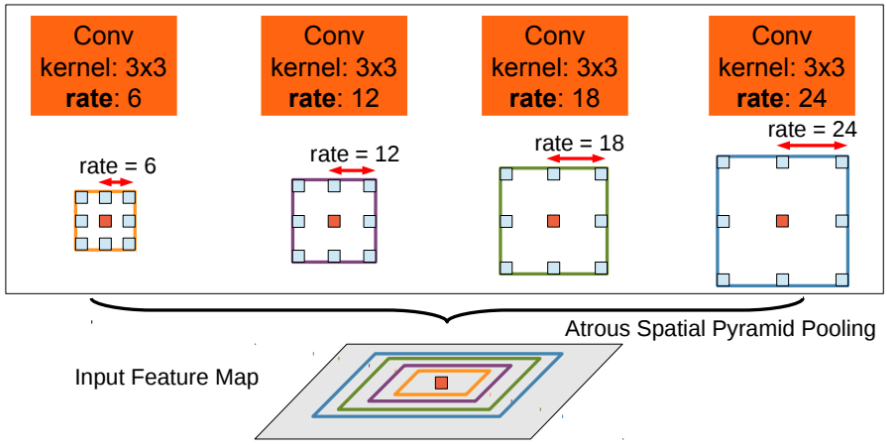


Figure 2.7: The illustration of 2D dilated convolution used in DeepLab. Image courtesy to [26]

For the FCN-8s, $2\times$ upsampling layer, i.e., deconvolution layer with its stride of two, is applied on conv7, and sum up its output with pool4. Then, their summing output is further upsampled by a deconvolution layer with its stride of two and added to pool3. Finally, a deconvolution layer with its stride of eight is used to upsample and produce the final segmentation map so that the pixel-wise prediction can be realized by a one-pass inference. From now, the research of pixel-wise prediction started to enter a new area where FCNs were extensively explored.

2.1.7 DeepLab

Deeplab was first proposed for semantic image segmentation task [26], and then, it has been evaluated to the second and third versions. The key contribution of DeepLab models is the dilated convolution (or atrous convolution). Figure 2.7 shows the illustration of 2D dilated convolution with different dilation rates. As can be seen, when apply the dilation rate as r on the $k \times k$ convolution kernel, its size will be enlarged to $k + (k-1)(r-1)$

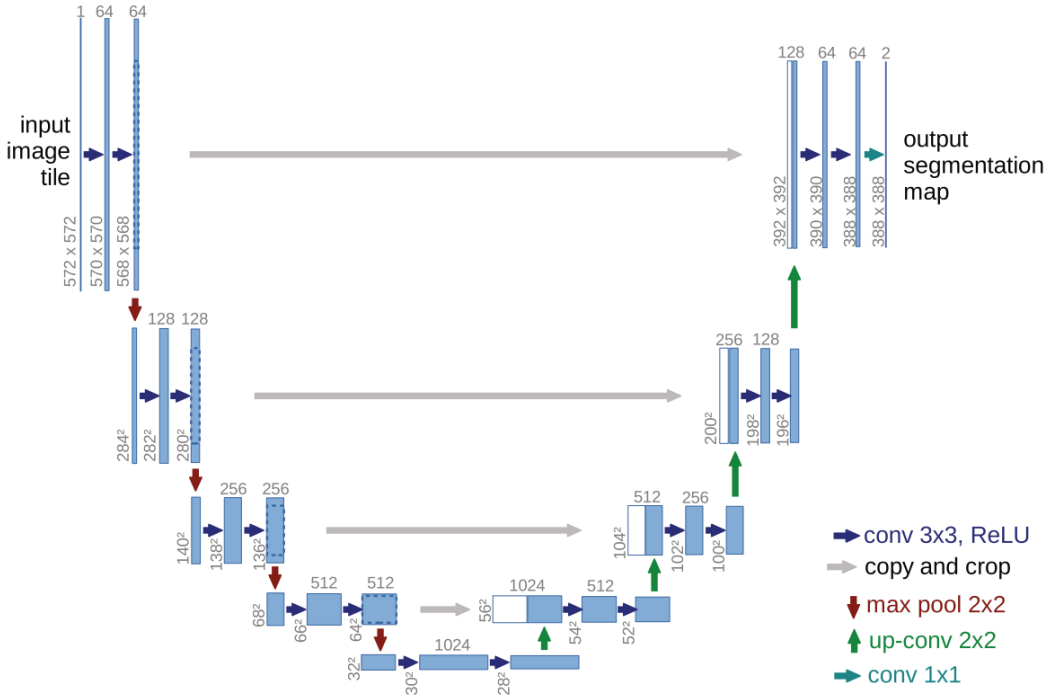


Figure 2.8: The architecture of Unet. Image courtesy to [8]

without any additional computational costs. Thus, the dilated convolution can increase the receptive fields by introducing larger dilation rates. Besides, with the dilated convolution, the sparse size of input features can be unchanged, which is very suitable for pixel-wise prediction tasks.

2.1.8 Unet

Unet was proposed to apply segmentation on 2D medical images and succeed as the first place in two challenges. It derives from the extension of FCN model [8]. Therefore, it also only uses the Conv layers to capture both of the local content and global contextual features from input images and estimate their pixel-wise predictions. As illustrated in Figure 2.8, Unet can be divided into two paths, the left downsampling (or contracting/encoder) path and the right upsampling (or expanding/decoder) path. In the downsampling path, after taking in the input image, it abstracts the image feature representations by five blocks. Each block consists of two Conv layers, a ReLU layer, and a max-pooling layer. Its upsampling path has the symmetric structure with the left path, including a deconvolution layer and two Conv layers with ReLU operations. The most important part in Unet is that it bridges the left and right paths with multiple skip connections, which transmit the feature maps from the left to concatenate with the features from the right. This delicate design not only enforces the model to understand input images with different receptive fields but also mitigates the gradient vanishing issue commonly seen in training deep learning models. Thus, Unet can reveal the multi-scale visual clues as the

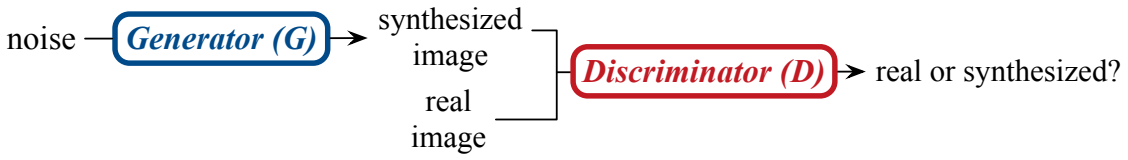


Figure 2.9: The illustration of vanilla GANs.

hierarchical feature representations by efficiently learning a pixel-prediction mapping.

2.2 Generative Adversarial Networks (GANs)

The basic structure of generative adversarial networks (GANs) consists of a generator and a discriminator. Despite the specific architectures in its generator and discriminator, they work to compete with each other. The generator tries to synthesize the data that cannot be recognized by the discriminator, while the discriminator aims to accurately classify the synthesized data and its corresponding real data into *fake* and *real* labels, respectively. With their adversarial competition between them, the generator will synthesize more realistic data. Based on this basic concept, four typical GANs with their learning approaches will be reviewed as follows.

2.2.1 Vanilla GANs

The vanilla GANs was proposed to synthesize generic images in 2014 [27]. It is a new learning mechanism rather than a specific model. It could be built upon any architectures. For computer vision tasks, since CNNs show the dominating performance, this thesis only talks about CNNs based GANs. All the aforementioned CNNs only have one network to process images for prediction. Differently, a GANs contains two networks, i.e., a generator G and a discriminator D, which are trained together by adversarial learning mechanism. To be more specific, the vanilla GANs, as illustrated in Figure 2.9, tries to learn a synthesis mapping from random noise $\mathbf{z} \sim p_{noise}(\mathbf{z})$ to the target images $\mathbf{x}$ following the real image distribution p_{data}. To train its G and D, the loss function is formulated as:

$$\min_G \max_D \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{noise}(\mathbf{z})}[\log(1 - D(G(\mathbf{x})))], \tag{2.2}$$

where $G(\cdot)$ and $D(\cdot)$ indicate the outputs of G and D, respectively, and $\mathbb{E}$ means mathematical expectation. Through using this adversarial loss, the competition between the generator and discriminator will reach a Nash equilibrium, so that the generator can synthesize less blurred target images.

To enhance the training stability of GANs, Wasserstein GAN with gradient penalty (WGAN-GP) was proposed in [28]. It applies Wasserstein distance that has a smoother

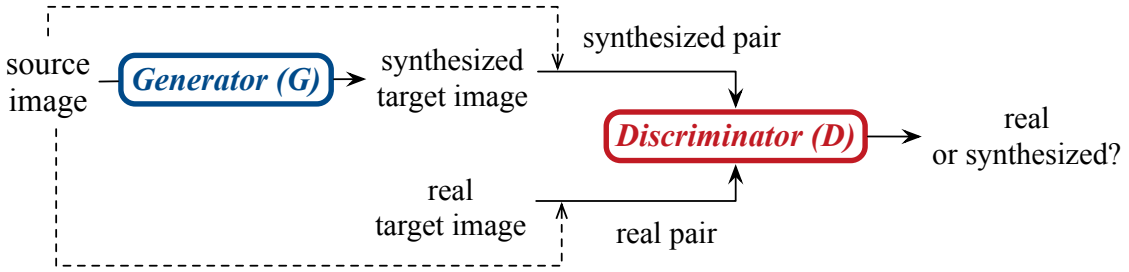


Figure 2.10: The illustration of Pix2pix.

gradient to adversarial learning and gradient penalty to enforce the constraint during training. It has demonstrated faster convergence and better performance in noise-to-image synthesis tasks.

2.2.2 Pix2pix

To add more constraints in the adversarial learning, conditional GANs (cGANs) was proposed in [29]. It conditions the learning process on a particular input data rather than only using the random noise, which offers an auxiliary guide for synthesis. In the paired-image synthesis, the input auxiliary information is a given source image $\mathbf{x} \sim p_{data}(\mathbf{x})$, and its corresponding target image is $\mathbf{y} \sim p_{data}(\mathbf{y})$. Therefore, the GANs for image-to-image synthesis is to learn the mapping between each pair of $\mathbf{x}$ and $\mathbf{y}$.

For the generic image-to-image synthesis, Pix2pix [30] demonstrates itself with the promising performance. Its structure is illustrated in Figure 2.10. Its generator has a Unet-like architecture to seize the hierarchical representations from an input source image $\mathbf{x}$, and learns to synthesize its target image $G(\mathbf{x})$ that resembles the corresponding real target image $\mathbf{y}$. Meanwhile, the CNN based discriminator D of Pix2pix tries to distinguish between the real image pair $(\mathbf{x}, \mathbf{y})$ and the synthesized image pair $(\mathbf{x}, G(\mathbf{x}))$. Benefit from the competition between G and D, the Pix2pix GANs can be trained to promote its synthesis ability. To realize the adversarial learning, during the training phase, the generator G learns to minimize the following objective:

$$\mathscr{L}_{GAN}^{G} = \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log(1 - D(\mathbf{x}, G(\mathbf{x})))] + \lambda_{l1}\mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{data}(\mathbf{x},\mathbf{y})}[\|y - G(\mathbf{x})\|_1]. \tag{2.3}$$

According to Equation 2.3, its first term tries to train the generator G to synthesize a realistic image $G(\mathbf{x})$ which could be misclassified by the discriminator D. At the same time, its second term applies an L1-norm penalty for the generator G to reduce the pixel-wise intensity distance between the real image $\mathbf{y}$ and the synthesized image $G(\mathbf{x})$. To balance these two terms, a hyper-parameter λ_{l1} is also employed in Equation 2.3.

The discriminator D in Pix2pix follows the formulated loss function:

$$\mathscr{L}_{GAN}^{D} = -\mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{data}(\mathbf{x},y)}[\log D(\mathbf{x}, \mathbf{y})] \; - \; \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log(1 - D(\mathbf{x}, G(\mathbf{x})))]. \tag{2.4}$$

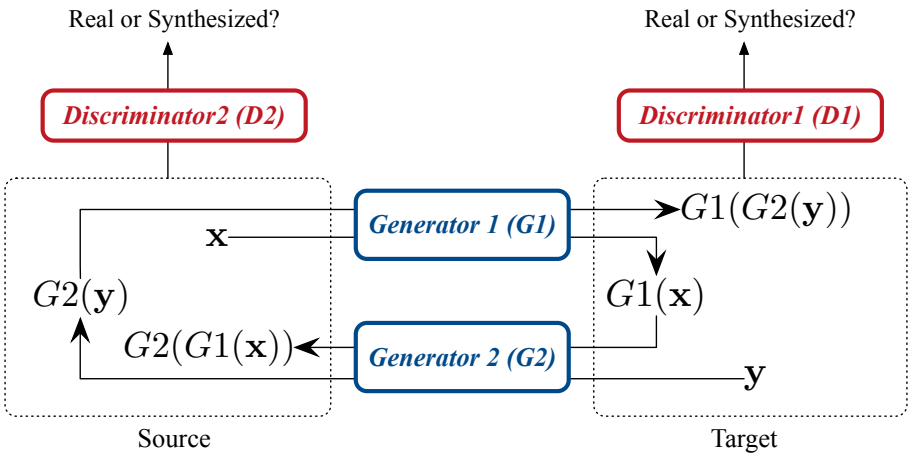


Figure 2.11: The illustration of cycleGAN.

Different from the training of G, Equation 2.4 trains the discriminator D to predict the correct image labels, i.e., zero for the synthesized image pair and one for the real image pair, in this binary classification.

After each forward-propagation during training Pix2pix, its generator G and discriminator D are updated in an iterative way conforming Equation 2.3 and Equation 2.4, respectively. Due to the battle between G and D, the synthesis ability of G could be further improved. When the entire GANs model completes training, only its generator would be applied to synthesizing target images.

2.2.3 CycleGAN

After Pix2pix, an unpaired image synthesis GANs called cycleGAN was proposed [31]. It includes two generators $G1$ and $G2$ and two discriminators $D1$ and $D2$ to conduct adversarial learning in two directions. The learning structure of cycleGAN is illustrated in Figure 2.11. Here, the image $\mathbf{x}$ from the source distribution and the image $\mathbf{y}$ from the target distribution are unpaired. During training, each image will experience a cycle generation. Fox example, $\mathbf{x}$ is first as the input of $G1$ to produce $G1(\mathbf{x})$, then $G2$ takes in $G1(\mathbf{x})$ to generate $G2(G1(\mathbf{x}))$. Thus, for $\mathbf{y}$, it also has two synthesized images $G2(\mathbf{y})$ and $G1(G2(\mathbf{y}))$ targeting at different image distributions. The two discriminators $D1$ and $D2$ correspondingly differentiate the synthesized images for these two distributions and also their real images. The adversarial losses in cycleGAN can be written as follows:

$$\mathscr{L}_{cycleGAN}^{G1,D1} = \mathbb{E}_{\mathbf{y}\sim p_{data}(\mathbf{y})}[\log D1(\mathbf{y})] + \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log(1 - D1(G1(\mathbf{x})))], \tag{2.5}$$

and

$$\mathscr{L}_{cycleGAN}^{G2,D2} = \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log D2(\mathbf{x})] + \mathbb{E}_{\mathbf{y}\sim p_{data}(\mathbf{y})}[\log(1 - D2(G2(\mathbf{y})))]. \tag{2.6}$$

Also, to ensure the cycle consistency, a cycle consistency loss is formulated as:

$$\mathscr{L}^{cyc}_{cycleGAN} = \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\|x - G2(G1(\mathbf{x}))\|_1] + \mathbb{E}_{\mathbf{y}\sim p_{data}(\mathbf{y})}[\|x - G1(G2(\mathbf{y}))\|_1]. \quad (2.7)$$

Therefore, the final objective function of cycleGAN is:

$$\mathscr{L}^{final}_{cycleGAN} = \mathscr{L}^{G1,D1}_{cycleGAN} + \mathscr{L}^{G2,D2}_{cycleGAN} + \lambda \mathscr{L}^{cyc}_{cycleGAN}, \quad (2.8)$$

the hyper-parameter λ is to balance these three terms.

2.2.4 Perceptual GANs

The perceptual GANs was proposed to involve the perceptual information of images into the adversarial learning for generic image synthesis [32]. It introduced a new perceptual discriminator which can be inserted into the existing GANs. The perceptual discriminator additionally uses a VGG encoder that is pretrained on the big generic image dataset ImageNet to extract the features of the real and the synthesized sample pairs and enforce their features to be similar through adversarial training. In this way, the perceptual GANs can synthesize the images with more perceptual information by its generator.

2.3 Deep CNNs based Per-voxel Regression on Medical Images

Per-voxel regression on medical images, i.e., medical image synthesis, is defined as a mapping between the unknown target-modality images and the given source-modality images. Current approaches can be roughly grouped into two categories. The first category refers to atlas-based methods. These methods utilize the paired image atlases of source- and target-modalities to calculate the atlas-to-image transformation in source-modality, and then explore this transformation to synthesize target-modality-like images from their corresponding target-modality atlases [33–37]. Since most atlases are built upon healthy subjects, these methods perform less satisfactorily on the images with pronounced abnormalities. The second category, learning-based methods, can mitigate this issue. Specifically, these methods directly learn a mapping from source-modality to target-modality. Once a training set appropriately contains pathology, such information could be captured by the learned model, so that abnormalities, such as brain tumors, can also be synthesized in target-modality images.

A large category of the learning-based synthesis methods train a nonlinear model that maps each small source-modality patch to the voxel at the center of the corresponding patch having the same location in target-modality [38–40]. Meanwhile, all these

mentioned patch-based methods have a limitation that the important spatial relationship among the small patches in the same image are ignored, leading to contrast inconsistency in the synthesized image. To alleviate this issue, global spatial information is additionally captured by multi-resolution patch regression in [41] for cross-modality image synthesis. However, the handcrafted features used in the above methods [38–41] have limited descriptive power to represent the complicated contextual details in images, which in turn affects synthesis quality. Moreover, in these methods, patch-based estimation is usually applied to each individual voxel, and the final estimation of a whole image is determined by a large number of highly overlapped patches. Therefore, such methods usually lead to over-smoothed synthesized images, and incur heavy computational cost.

To deal with the above problems, deep learning based models, especially CNNs, have been used to automatically learn features with better descriptive power [7, 24]. This section mainly focuses on discussing the CNNs based synthesis of three important medical images, i.e., CT, PET, and MRI. According to the imaging modalities of the source and target images, the applications of medical image synthesis can be roughly categorized into two classes. The first one is within-modality synthesis, which targets to predict the higher-quality images from the given source images of the same modality with lower quality. It includes the synthesis of full-dose CT from the low-dose [42–44], 7T MR images from the 3T [45–47], and full-dose PET images from the low-dose [48]. The second class is cross-modality synthesis. It aims at extracting the visual information from the input source-modality and transforming to generate the target-modality images. It usually consists of the synthesis between MR and CT images [47, 49–56], CT and PET images [57–59], MR and PET image [60, 61], and various MR modalities, like T1, T2, fluid-attenuated inversion recovery (FLAIR), and magnetic resonance angiography (MRA) [62–67]. Despite their different applications, they share the same technical essence that builds a mapping from the source to the target images.

To learn the medical image synthesis mapping, deep CNNs are the most prevalent choice in the recent years because of their successful applications in computer vision tasks. These deep CNNs models can be classified into two groups, i.e., conventional CNNs and CNNs based GANs. Here, the word “conventional” is used to show that these CNNs models do not have discriminators rather than that they are out-of-date. They also have developed to various advanced structures. Table 2.1 collects the key works using conventional CNNs for medical image synthesis. The first attempt is deep location-sensitive network which was proposed in [62]. Although it applies multiplicative interactions to extracting spatial information from the input images, which is different from the common CNNs using spatial pooling, it is still reviewed here due to its used convolution operation. It crops $3 \times 3 \times 3$ small patches from the input images, and separately estimates the target intensity values of their centered voxels. With sliding windows, the whole target images can be synthesized. A custom three-layer CNNs was also proposed

Table 2.1: Conventional CNNs based medical image synthesis publications.

METHOD	PUBLICATION	DATASET	OBJECT	TASK
location-sensitive network	Van et al. [62]	NAMIC [68]	brain	cross-modality MR
custom three layer-CNNs	Chen et al. [42]	NBIA [69]	multiple organs	full-dose CT
FCNs	Nie et al. [49]	-	pelvic	MR-to-CT
residual encoder-decoder	Chen et al. [43]	NBIA [69]	multiple organs	full-dose CT
autoencoder	Liu et al. [52]	-	brain	MR-to-CT
encoder-decoder	Chartsias et al. [63]	ISLES2015 [16] BRATS2015 [14] IXI [15]	brain	cross-modality MR
Unet	Han et al. [50]	-	brain	MR-to-CT
Unet	Leynes et al. [51]	-	pelvic	MR-to-CT
residual CNNs	Chaudhari et al. [46]	OAI [70]	knee	3T-to-7T MR
residual CNNs	Zend et al. [45]	Brainweb [71] NAMIC [68]	brain	3T-to-7T MR
cascade CNNs	Xiang et al. [48]	-	brain	full-dose PET
CNNs with reconstruction	Zhou et al. [72]	BRATS2018[19]	brain	cross-modality MR
CNNs in wavelet domain	Kang et al. [44]	low-dose CT [73]	head, chest and abdomen	full-dose CT

using the same learning strategy [42]. They outperformed many previous handcrafted features applied methods. However, this patch-to-pixel/voxel prediction is low efficiency. After FCNs was popular in pixel-wise generic image prediction tasks, it was extended in MR-to-CT synthesis work [49]. By using the fully convolution structure, this model realizes the patch-to-patch learning and improves the synthesis efficiency. Based on FCNs, different encoder-decoder and autoencoder architectures [43, 52, 63] were designed and successfully applied to producing the same-size output. Moreover, to acquire the multi-scale feature representations and mitigate the gradient vanishing issue, skip connections were added into encoder-decoder structures as Unet-like models that conduct the global-level synthesis [50, 51]. In addition, inspired by the shortcut connection in ResNet, residual learning was integrated into CNNs to handle the saturated accuracy caused by some structures and further improved the synthesis performance [43, 45, 46]. The work in [48] attempted to exploit cascade CNNs, which has a three-stage structure. For its second and third stages, the initial input of the entire CNNs and the output from last stage are combined as the new input. In this way, the synthesis results can be polished. Whereas, it needs more computational costs than the single model, and the improvement is limited after stacking more stages. The CNNs proposed in [19] created an additional reconstruction path for each source modality to cooperate with more regularization during feature extraction for synthesis. Different from the above CNNs directly processing input images/patches, the model in [44] takes in the image features from wavelet domain. It is especially suitable for within-modality synthesis due to that it may actively reduce the noise in wavelet domain shared by all the same-modality images.

Advanced GANs is the second group of deep CNNs popularly applied in medical image synthesis. It has an additional discriminator to differentiate the synthesized images estimated from its generator and the corresponding target ground-truths. The related key GANs based publications are collected and introduced in Table 2.2. Since medical image synthesis tasks need to condition on the given source images, the image-to-image translation by cGANs is a more appropriate choice than the noise-to-image synthesis by vanilla GANs. Most cGANs [57, 60, 64, 66] for medical image synthesis derive from Pix2pix [30] due to its high synthesis quality on generic images. Their Unet-like generators perform well to apply the global-level synthesis, while their discriminators work as PatchGAN classifiers to compete with the generators. With the PatchGAN architecture, the discriminators can distinguish the image patches rather than the whole images are real or fake, which helps to consider more style and texture details in images during the classification. After their generators and discriminators come to Nash equilibrium, the cGANs are well trained and ready for synthesis. The generators in cGANs can be replaced by other conventional CNNs. Both of the works [54] and [65] incorporated ResNet based generators into the adversarial learning, but no significant improvement was brought up by the replacement of generators. Besides, some cGANs involve more annotated object

Table 2.2: GANs based medical image synthesis publications.

METHOD	PUBLICATION	DATASET	OBJECT	TASK
FCN-cGAN (Pix2pix)	Ben et al. [57]	-	liver	CT-to-PET
cGANs (Pix2pix)	Yang et al. [66]	BRATS2015 [14]	brain	cross-modality MR
cGAN (Pix2pix)	Choi et al. [60]	ADNI [74]	brain	PET-to-MR
cGANs (Pix2pix)	Dar et al. [64]	MIDAS [75] BRATS2015 [14] IXI [15]	brain	cross-modality MR
cGANs (ResNet)	Emami et al. [54]	-	brain	MR-to-CT
cGANs (ResNet)	Olut et al. [65]	IXI [15]	brain	cross-modality MR
cGAN with tumor label input	Bi et al. [58]	-	thorax	CT-to-PET
cGANs (two *G*s and four *D*s)	Mok et al. [76]	BRATS2015 [14]	brain	label-to-MR
cGANs with CasNet generator	Armanious et al. [59]	-	brain	PET-to-CT, MR correction, and PET denosing
cascade cGANs	Wei et al. [61]	-	brain	MR-PET
cascade cGANs with gradient loss	Nie et al. [47]	ADNI [74]	brain and pelvic	3T-to-7T MR and MR-to-CT
cycleGAN	Chartsias et al. [53]	MM-WHS [77]	cardiac	CT-to-MR
cycleGAN and UNIT	Welander et al. [67]	Human Connectome [78]	brain	cross-modality MR
cycleGAN with segmentors	Zhang et al. [56]	-	cardiac	MR-CT
cycleGAN with gradient related loss	Hiasa et al. [55]	-	musculoskeletal	MR-to-CT

information into the synthesis learning. The cGAN model in [58] added the tumor labels and the source images together as the input of its generator to further condition the synthesis on different pathological cases. The work [76] directly synthesizes the target MR images from the tumor labels without any source images. Because of the higher difficulty of label-to-image synthesis, its cGANs consists of two generators and four discriminators. The two generators conduct coarse and fine synthesis, respectively, and its four discriminators operate the discrimination at four image scales to enforce more constraints. Also, similar to the conventional CNNs, cascade structure is exploited in cGANs to polish the synthesis. Armanious et al. [59] introduced cascade Unet-like blocks in its generator, and cascade cGANs were proposed in [47, 61] to promote the synthesis performance. In addition, other works [53, 55, 56, 67] followed cycleGAN [31] in medical image synthesis. The structure of cycleGAN is built by two generators and two discriminators to conduct unpaired generic image translation through the adversarial losses and reconstruction consistency loss. Among these medical image methods, the models in [53, 55, 56] are also applied on unpaired synthesis. To give more control on these mappings, segmentation labels are additionally exploited in [53, 56] so that the synthesized images contain more pathological information. In the work [67], cycleGAN is modified to synthesize paired images. Since its GANs needs two generators to complete the reconstruction cycle, more computational sources are in demand during training. Moreover, image gradient related losses are imposed in the training of GANs [47, 55], which helps their generators more sensitive to the subtle changes of objects in the input medical images.

After reviewing the above deep CNNs based medical image synthesis approaches, it can be summarized that GANs not only takes the advantages of flexible CNNs architectures in its generator to synthesize target images by seizing the essential features from source images but also benefits from the adversarial mechanism that further enforce the synthesized images resembling the real images. All the above CNNs based models have not handled the data variation issue during the learning of their unified CNNs models. Therefore, how to build a more effective GANs for cross-modality MR image synthesis and how to further cope with the data variation to improve the synthesis performance of GANs will be explored in the following chapters.

2.4 Deep CNNs based Per-voxel Classification on Medical Images

Medical image segmentation identifies the interested regions and delineates their contours in medical images. Since the segmentation results delivery the crucial information about the shape and volume related clinical parameters, the more effective and efficient automatic segmentation approaches are always under high demand. Recently, the deep CNNs

models have been widely explored for medical image segmentation due to their advanced performance. These deep CNNs are applied to two categories of segmentation tasks, i.e., organ or substructure segmentation and lesion segmentation.

In the organ or substructure segmentation, the target objects could be pancreas, liver, prostate, kidneys, brain tissues, ventricles in cardiac images, and other substructures [5]. The location and appearance information of these objects assists the following computer-aided diagnosis and treatment planning. This category of medical image segmentation benefits from the similar locations of objects varying with the scanned images from different patients, especially in the registered images, but still requires the efforts to accurately depict the boundaries of objects. Tables 2.3 and 2.4 review the key relevant publications for organ or substructure segmentation. As can be seen, the development of using deep CNNs in this segmentation application started from introducing the classification models into the pixel-/voxel-wise labelling [79, 80, 82–85, 87–89]. These models conducted patch-to-pixel/voxel segmentation. Each cropped patch from a given image is taken as the input. After it is processed by convolutional and fully connected layers, the CNNs will estimate the segmentation label of its center pixel/voxel. By moving the sliding windows over the entire image, the final segmentation results can be predicted. This approach is first applied by 2D CNNs to the axial slices of 3D medical images [79, 80, 82–85]. Then, 2.5D CNNs were proposed to extract the features from all the three views of medical images [87–89]. The axial, coronal, and sagittal patch slices that have the same center voxel are separately put into the three input pathways of these CNNs, and then fuse their deep feature maps to estimate the label. With this strategy, the label annotation of each voxel can see the three views of its surroundings, so that more volumetric information will be captured, and the segmentation accuracy will be improved. Besides, some effective approaches that were demonstrated in generic image segmentation, like multi-scale feature extraction, residual learning, and cascade structure, were integrated into some of the above CNNs to further promote their segmentation performance [83–85, 89]. Whereas, directly applying the classification models to segmentation task results is in low inference efficiency. The insufficient global-level information acquisition in these classification models also lowers their segmentation performance. Therefore, FCNs that succeeded in generating the same-size labels on the given generic images began to explore its applications in medical image segmentation. It has been extended to different 2D, 2.5D, and 3D variants in various tasks [91, 93–95]. Among them, to better delineate the small objects, like pancreas, in relatively large images, bounding box was brought into the segmentation task to realize coarse-to-fine estimation. Also, deep supervision and residual learning strategies were used in [94, 95] to build more back-propagation paths during training. In addition, following the successful Unet [8], CNNs with different Unet-like structures were proposed for global-level segmentation [96–98, 102]. Some of them not only took the advantages of both local and global feature representations from

Table 2.3: Deep CNNs based organ or substructure segmentation publications: Part 1.

METHOD	PUBLICATION	DATASET	OBJECT	MODALITY
2D CNNs for patch classification	Farag et al. [79]	-	pancreas	CT
2D CNNs for patch classification	Li et al. [80]	SLiver07 [81]	liver	CT
2D CNNs for patch classification	Thong et al. [82]	-	kidneys	CT
2D multi-scale CNNs for patch classification	Moeskops et al. [83]	-	brain	MR
2D multi-scale CNNs for multi-task patch classification	Moeskops et al. [84]	-	brain, breast, and cardiac	MR
2D cascade multi-scale CNNs with residual learning for patch classification	Chen et al. [85]	MRBrainS [86]	brain	MR
2.5D CNNs for patch classification	Roth et al. [87]	NIH-CT [87]	pancreas	CT
2.5D CNNs for patch classification	Zreik et al. [88]	-	cardiac	CT
2.5D multi-scale CNNs for patch classification	Brebisson et al. [89]	multi-atlas [90]	brain	MR
2D FCNs	Tran et al. [91]	LVSC [92]	cardiac	MR
2.5D FCNs with bounding box for coarse-to-fine prediction	Zhou et al. [93]	NIH-CT [87]	pancreas	CT
3D FCNs with deep supervision	Dou et al. [94]	SLiver07 [81]	liver	CT
3D FCNs with residual learning and deep supervision	Yu et al. [95]	PROMISE12 [96]	prostate	MR

Table 2.4: Deep CNNs based organ or substructure segmentation publications: Part 2.

METHOD	PUBLICATION	DATASET	OBJECT	MODALITY
recurrent Unet-like CNNs	Poudel et al. [97]	LVSC [92]	cardiac	MR
3D Unet with deep supervision	Zhu et al. [98]	-	prostate	MR
3D Vnet	Milletari et al. [99]	PROMISE12 [96]	prostate	MR
3D DenseNet with early fusion of multi-modalities	Dolz et al. [100]	iSEG2017 [101] and MRBrainS [86]	brain	MR
3D Unet-like GANs	Jia et al. [102]	PROMISE12 [96]	prostate	MR
3D GANs with deep supervision	Yang et al. [103]	SLiver07 [81]	liver	CT
3D GANs with dilated convolution	Moeskops et al. [104]	multi-atlas[90]	brain	MR

images, but also exploited the inter-slice spatial dependence on 3D medical images [97] and more connections among neighboring and distant feature maps [99] to increase the segmentation accuracy. Moreover, dense blocks and dilated convolutional layers were incorporated into the architectures of CNNs from [100] and [104], respectively. They enhanced the learning of hierarchical visual information for segmentation. Some CNNs based GANs were also attempted [102–104]. Their generators produce the label maps which are classified by their discriminators during training and get more realistic label maps in their synthesis tasks.

The second category, i.e. lesion segmentation, consists of the tasks, such as liver lesion, brain lesion, and brain tumor segmentation. Compared with the organ or substructure segmentation, this application is more challenging since the target lesions always have the arbitrary locations and various appearance varying from patient to patient. Thus, hierarchically feature representations including the pathological information about local content and global context should be learnt. Table 2.5 lists the key deep CNNs based lesion segmentation publications. In this segmentation application, the CNNs models also experienced the development from using fully connected layers [6, 105] for patch classification to employing FCNs in directly predicting the 2D/3D labels for given patches [107, 108]. As a FCNs, the model, called DeepMedic, proposed in [107] won the first place in

Table 2.5: Deep CNNs based lesion segmentation publications.

METHOD	PUBLICATION	DATASET	OBJECT	MODALITY
2D multi-scale CNNs for patch classification	Havaei et al. [6]	BRATS2013 [14]	brain tumor	MR
3D cascade CNNs for patch classification	Valverde et al. [105]	ISBI-MS [106]	brain lesion	MR
3D multi-scale FCNs with CRFs	Kamnitsas et al. [107]	ISLES2015 [16] and BRATS2015 [14]	brain lesion and tumor	MR
2.5D FCNs with parallel learning for multi-modalities	Roy et al. [108]	ISBI-MS [106]	brain lesion	MR
2D and 3D DenseUnet with hybrid feature fusion	Yang et al. [103]	LiTS [109]	liver lesion	CT
2D cascade Unets with CRFs	Christ et al. [110]	-	liver lesion	CT
3D Unet	Nair et al. [111]	-	brain lesion	MR
3D Unet with residual learning	Isensee et al. [17]	BRATS2017 [19]	brain tumor	MR
3D Unet with residual learning	Isensee et al. [112]	BRATS2018 [19]	brain tumor	MR
3D Unet with reconstruction branch	Myronenko et al. [113]	BRATS2018 [19]	brain tumor	MR
3D Unet with three-branch units	Chen et al. [114]	BRATS2018 [19]	brain tumor	MR
3D Unet with dilated multi-fiber units	Chen et al. [18]	BRATS2018 [19]	brain tumor	MR
3D ensemble Unet and FCNs	Kamnitsas et al. [115]	BRATS2017 [19]	brain tumor	MR
3D ensemble CNNs with attention	Zhou et al. [116]	BRATS2018 [19]	brain tumor	MR

BRATS2015 challenge [107]. It has two feature extraction paths to process the two-scale input patches and fuses their features to label their shared centered part. Since its inputs are still the local patches, a fully connected CRFs was applied to help the labels for the whole images more consistent. Whereas, the CRFs is a separate postprocessing step from the deep learning, which leads that the improvement brought from the CRFs is not significant. Thus, the CNNs for global-level segmentation have been more explored. As mentioned, the Unet-like CNNs can capture both the local and global information together in a single model, which is especially beneficial to locating the lesions with arbitrary shapes. Therefore, recently, most outstanding works are based on Unet [17, 18, 111–115]. Among them, Isensee et al. [17] build more connections among feature maps with residual learning as modified 3D Unet with 26 convolutional layers to enhance the pathological information extraction, which achieved the third place in BRATS2017 [19]. Its architecture was also widened as No New-Net using more convolution kernels in each layer and won the second place in BRATS2018 [19]. The first-place winner in BRATS2018 [19] is the CNNs model called NVDLMED [113]. It cooperated an additional image reconstruction branch with the common decoding path in Unet to learn more generalized object content in its encoding path for the final segmentation. The Unet-like model that was called DMFNet from [18] attained the comparable tumor segmentation results with NVDLMED in BRATS2018 dataset but only used about 1/10 parameters of NVDLMED. DMFNet replaced the common convolutional layers with the new dilated multi-fiber units in 3D Unet to adaptively control its receptive field for the better segmentation. Also, S3S-Unet [114] performed well in BRATS2018. It brought a three-branch unit into 3D Unet to replace the traditional convolutional layers so that its trainable parameters could be largely reduced to increase the learning efficiency. Moreover, ensemble models were proposed in [115, 116]. The multiple CNNs in the large models can complement with each other and further enhance the generalization of learnt features. Kamnitsas et al. [115] integrated Unet, FCNs, and DeepMedic together into its ensemble model and won the first place in BRATS2017. Zhou et al. [116] additionally employed attention in the ensemble CNNs to further improve the performance. However, training the ensemble CNNs requires more computational resources due to their multiple models than a single model.

Overall, Unet-like CNNs that are able to conduct the global-level mapping have shown better medical image segmentation performance, especially for the lesion involved images. However, all these CNNs have not addressed the significant visual variation among all the images when learning their unified models. Thus, how to effectively mitigate the variation issue and improve the performance of the unified Unet-like CNNs models for brain tumor segmentation will be discussed in this thesis.

2.5 Evaluation Metrics

2.5.1 Evaluation Metrics for Medical Image Synthesis

In the literature, three evaluation metrics are commonly used to measure the medical image synthesis results [39, 63]. They are peak signal-to-noise ratio (PSNR), normalized mean squared error (NMSE), and structural similarity index (SSIM) [117]. If denote the real target image and its corresponding synthesized target image as $\mathbf{y}$ and $\hat{\mathbf{y}}$, respectively, PSNR and NMSE are used to measure the absolute differences between $\mathbf{y}$ and $\hat{\mathbf{y}}$ while SSIM evaluates the image degradation from $\mathbf{y}$ to $\hat{\mathbf{y}}$ through calculating the local changes of the perceived structural information.

PSNR is computed as follows:

$$\mathrm{PSNR}(\mathbf{y},\hat{\mathbf{y}}) = 10\log_{10}\frac{MAX^2_{range}(\mathbf{y},\hat{\mathbf{y}})}{N^{-1}_{voxel}\|\mathbf{y}-\hat{\mathbf{y}}\|_2^2}, \tag{2.9}$$

where the symbol N_{voxel} means the entire voxel number in $\mathbf{y}$ or $\hat{\mathbf{y}}$, and $MAX_{range}(\mathbf{y},\hat{\mathbf{y}})$ indicates the maximum intensity range of $\mathbf{y}$ and $\hat{\mathbf{y}}$. Thus, the synthesis accuracy can be measured by PSNR in the logarithmic axes. Its higher resulting value shows the higher synthesis quality of $\hat{\mathbf{y}}$.

NMSE can be calculated as follows:

$$\mathrm{NMSE}(\mathbf{y},\hat{\mathbf{y}}) = \frac{\|\mathbf{y}-\hat{\mathbf{y}}\|_2^2}{\|\mathbf{y}\|_2^2}. \tag{2.10}$$

The NMSE metric is applied to measure the voxel-wise distance between the intensities of $\mathbf{y}$ and $\hat{\mathbf{y}}$. Its lower value indicates the better synthesis.

SSIM is defined in the following:

$$\mathrm{SSIM}(\mathbf{y},\hat{\mathbf{y}}) = \frac{(2\mu_{\mathbf{y}}\mu_{\hat{\mathbf{y}}}+c_1)(2\sigma_{\mathbf{y}\hat{\mathbf{y}}}+c_2)}{(\mu^2_{\mathbf{y}}+\mu^2_{\hat{\mathbf{y}}}+c_1)(\sigma^2_{\mathbf{y}}+\sigma_{\hat{\mathbf{y}}}\hat{\mathbf{y}}^2+c_2)}, \tag{2.11}$$

where the symbols $\mu_{\mathbf{y}}$, $\mu_{\hat{\mathbf{y}}}$, $\sigma_{\mathbf{y}}$, and $\sigma_{\hat{\mathbf{y}}}$ denote the means and variances of image $\mathbf{y}$ and $\hat{\mathbf{y}}$, respectively, and $\sigma_{\mathbf{y}\hat{\mathbf{y}}}$ means the covariance of $\mathbf{y}$ and $\hat{\mathbf{y}}$. Higher SSIM values validate the better synthesis.

Following the literature, the above three measures will be used in this thesis to evaluate the synthesis performance of the proposed GANs models.

2.5.2 Evaluation Metrics for Medical Image Segmentation

Most works in literature [14] employed two measures, i.e, Dice score and Hausdorff distance, to evaluate their segmentation performance. Dice score can show the overlapped

regions between the groundtruth and the segmented results, while Hausdorff distance can compute the surface distance between their segmentation boundaries. Here, the binary groundtruth map of a target object is denoted by $T \in \{0,1\}$, and the predicted segmentation map of the object is $P \in \{0,1\}$. The value 1 means that this voxel belongs to this object in the map. In contrast, 0 denotes that the voxel is in the background. To better give the definitions of the two metrics, T_1 and P_1 are used to represent the sets of voxels where $T = 1$ and $P = 1$, respectively.

The Dice score metric is calculated as:

$$\text{Dice}(T,P) = \frac{2|T_1 \wedge P_1|}{(|T_1| + |P_1|)}, \tag{2.12}$$

where $|\cdot|$ indicates the number of voxels in its inner set, and symbol $\wedge$ denotes the logical operator *AND*. The higher Dice score means more accurate segmentation of P.

The definition equation of Hausdorff distance is following:

$$\text{Hausdorff}(T,P) = \max\{\sup_{t \in \vartheta T_1} \inf_{p \in \vartheta P_1} d(t,p), \sup_{p \in \vartheta P_1} \inf_{t \in \vartheta T_1} d(p,t)\}, \tag{2.13}$$

where ϑT_1 and ϑP_1 are the surfaces of T_1 and P_1, respectively, and $d(t,p)$ is the least-squares distance between the point t and the point p, and vice versa. Besides, sup and inf correspondingly mean supremum and infimum. Thus, the lower Hausdorff distance indicates better segmentation.

In the following part of this thesis, the above two metrics will be employed to measure the segmentation performance.

Chapter 3

3D cGAN for Cross-modality MR Image Synthesis

The existing CNNs based GAN models have achieved promising performance in per-voxel regression tasks on medical images, such as cross-modality brain MR image synthesis. To meet the higher requirement of synthesis quality, more advanced and effective GANs that can seize the local details about the 3D nature of objects and the global image contextual information are still in high demand. This chapter develops a GAN model based on the 3D CNNs architecture to synthesize the target-modality MR images from their corresponding source-modality images. Moreover, a local adaptive mapping approach is explored to further polish the synthesized target-modality images from the GAN model.

3.1 Introduction

MRI, which can produce different modalities of images by setting task-specific scanning parameters, has been broadly exploited in medical image analysis [118, 119]. During the analysis, the images from multiple MR imaging modalities (e.g., T1-weighted, T2-weighted, and FLAIR) are processed together, since each modality shows unique soft tissue contrast. For example, multi-modality MR images have been collectively utilised to study the neuroanatomy of human brains for disease diagnosis [120] or therapy planning [121]. The complementary information from multi-modalities demonstrates better predictive power than that from a single imaging modality. Also, the benefits of using multi-modality MR images for brain lesion segmentation have been widely recognised [6, 107]. At the same time, due to modality missing and modality inconsistency between different clinical centers [63], the high demand of employing multiple MR imaging modalities for analysis is not always met in clinic and research, which adversely affects the quality of diagnosis and treatment. Therefore, cross-modality MR image synthesis

has recently aroused increasing research interest, and more and more investigations have been conducted to cope with the limitation of insufficient modalities in clinic and research [40, 41].

3.2 Motivation

Recently, conditional Generative Adversarial Network (cGAN) [29] has demonstrated itself to be a promising method for image synthesis. Different from traditional learning based methods, cGAN consists of two modules, a generator to learn the mapping for realistic images and a discriminator to distinguish the real and the synthesized images. By training the two modules to beat each other, cGAN has achieved excellent performance on medical image synthesis, as shown in [57, 122]. Nevertheless, these methods synthesize each individual slice independently along the axial direction and then concatenate them into a 3D image. This results in discontinuous estimation along the coronal and the sagittal directions. However, image information along these two directions are also crucial for medical image synthesis and analysis. Although some recent work [123] applies 3D cGAN to predict small patches to eliminate the discontinuity caused by 2D estimation, learning on small patches is insufficient to extract both local and global contextual relationship among voxels, and could hurt the synthesis of brain tumor images, as shown in section 3.4.

In this chapter, we propose a 3D cGAN model for MR image synthesis. It mitigates the problem of discontinuous estimation across slices caused by the 2D cGAN in the current literature. By considering large image patches and hierarchical features from skip connections, our 3D cGAN model could better synthesize MR images by taking contextual information into account. Moreover, to further improve the synthesized MR images for the segmentation task, a local adaptive synthesis method is proposed, which better depicts the local details of the synthesized MR images. Therefore, our method consists of both a global non-linear mapping and a local linear mapping from source-modality to target-modality MR images. The global non-linear mapping determines the similarity of synthesized images to real target-modality images at the whole image level, which is estimated by our proposed 3D cGAN model. The local linear mapping further improves the local details from source-modality images. We also evaluate the quality of synthesized MR images for an image recognition task, i.e., brain tumor segmentation. The synthesized target-modality images are utilized to help brain tumor segmentation from their corresponding source-modality images via training a CNNs that considers two imaging modalities jointly. The effectiveness of the proposed method is demonstrated on the public dataset 2015 Brain Tumor Segmentation Challenge (BRATS) [14].

3.3 Proposed Method

Our proposed cross-modality MR image synthesis framework generates target-modality-like images from source-modality images by a 3D cGAN model and a following local adaptive fusion to cater for the similarity at both whole image and local patch levels. The final synthesized target-modality images, together with source-modality images, are processed by a two-pathway 3D CNNs model to segment brain tumor. Its results can be also used to evaluate the synthesis quality of target-modality images.

3.3.1 3D cGAN

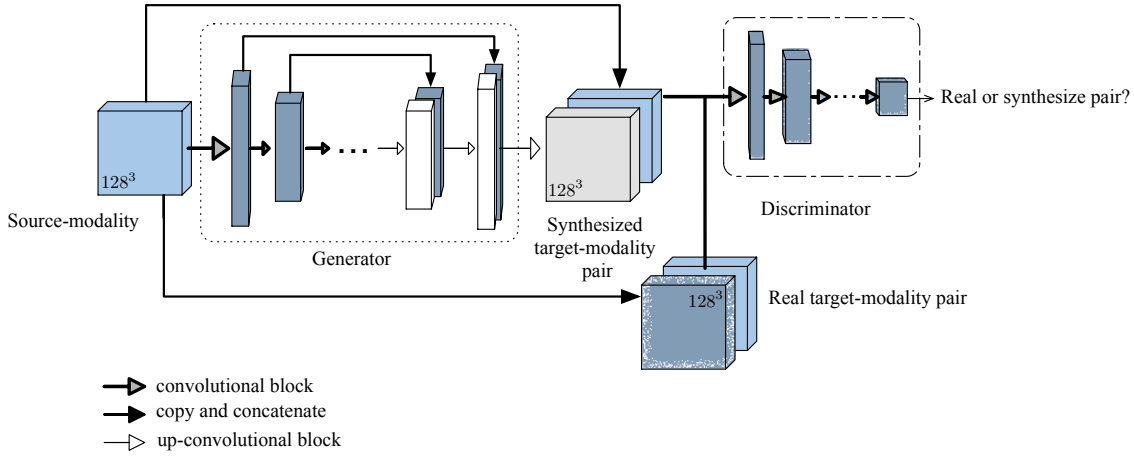


Figure 3.1: The proposed framework of 3D cGAN. It consists of a generator G and a discriminator D.

The framework of the proposed 3D cGAN is illustrated in Figure 3.1. Large patches ($128\times128\times128$), rather than a whole image ($240\times240\times155$) are used as the input of our model to deal with the limited number of training samples, as well as controlling the number of parameters to learn. The basic idea and detailed architecture of our 3D cGAN are presented as follows.

Basic Ideas

The original GAN [27] consists of two modules, the generator G and the discriminator D, struggling with each other to synthesize images $G(\mathbf{z})$ resembling real images $\mathbf{y}$ from the random vector $\mathbf{z}$ and meanwhile distinguish $\mathbf{y}$ from $G(\mathbf{z})$. cGAN has extended the original GAN to capture auxiliary information $\mathbf{x}$ in both the generator G and the discriminator D. For our cross-modality image synthesis, $\mathbf{x}$ is the source-modality MR images that are the input to synthesize target-modality-like images $G(\mathbf{y})$. The real pair $(\mathbf{x},\mathbf{y})$ and the estimated pair $(\mathbf{x},G(\mathbf{x}))$ are differentiated by the discriminator D. The generator G and

the discriminator D are trained simultaneously, as if they are following a two-player min-max game with the following objective:

$$\min_G \max_D \mathscr{L}_{cGAN}(G,D) = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{data}(\mathbf{x},\mathbf{y})}[\log D(\mathbf{x},\mathbf{y})] + \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log(1-D(\mathbf{x},G(\mathbf{x})))], \tag{3.1}$$

where $G(\cdot)$ and $D(\cdot)$ denote the outputs of the generator and the discriminator, respectively.

Furthermore, to ensure the voxel-wise similarity between the synthesized and the real images, an L1-norm penalty [30] is also utilized and formulated as follows:

$$\mathscr{L}_{L1}(G) = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{data}(\mathbf{x},\mathbf{y})}[\|\mathbf{x}-G(\mathbf{y})\|_1]. \tag{3.2}$$

Combining cGAN objective and L1 loss, the final objective function is formulated as:

$$\mathscr{L}_{total} = \arg\min_G \max_D \mathscr{L}_{cGAN}(G,D) + \lambda \mathscr{L}_{L1}(G), \tag{3.3}$$

where λ is a hyperparameter to balance the two terms.

Detailed Architecture

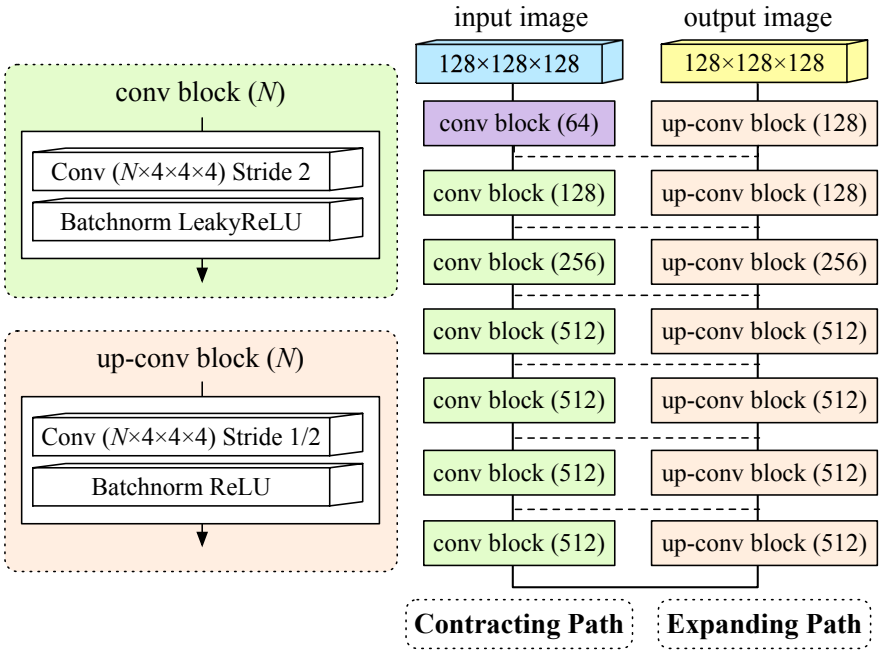


Figure 3.2: Generator architecture. All the convolutional and up-convolutional blocks contain convolutional, batch normalization, and ReLU layers. In addition to these three layers, drop-out is applied to the first three blocks on the expanding path. Batch normalization is not used in the first block of the contracting path. Dashed lines mean the skip connections between contracting and expanding paths by copy and concatenation. The slope of LeakyReLUs is 0.2.

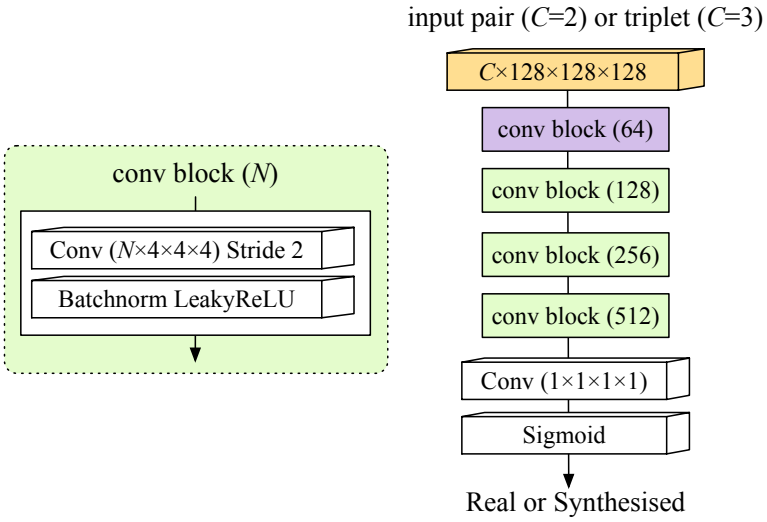


Figure 3.3: Discriminator architecture. Convolutional, and LeakyReLU layers with the slope of 0.2 are applied to all conv blocks. Batch normalization is not used in the first blocks of *D*.

Unet, as a CNNs-based model, has been proposed to analyze whole images or large image patches in the literature [8]. It acquires global contextual information from the input and ensures the spatial contiguity of the output. The typical characteristic of Unet architecture is the contracting and expanding paths with multiple skip connections between them. Using this structure, Unet can capture the hierarchical features of an input image, and mitigate the gradient vanishing caused by the long backpropagation when training deep networks [8]. It has been extended into 3D variants to better deal with 3D medical images [99, 124]. To benefit from the structure of Unet, we design the generator of Ea-GANs as a 3D Unet-like network. It is symmetric with seven convolutional (conv) blocks in its contracting path and seven up-convolutional (up-cov) blocks in its expanding path. Between each conv block and the corresponding up-conv block, skip connection is applied to capture multi-depth information of source-modality images effectively. The specific construction of this generator is shown in Figure 3.2. Coupled with a five-convolutional-layer 3D discriminator (including Sigmoid function), it constitutes our proposed 3D cGAN for target-modality image synthesis from source-modality. The detailed architecture of the proposed discriminator is illustrated in Figure 3.3.

3.3.2 Subject-specific Local Adaptive Fusion

In this chapter, our ultimate goal is to synthesize the target-modality images that could help boost the segmentation performance of brain tumor. This puts forward higher requirements on the quality of the output images compared with those synthesis methods merely focusing on improving PSNR (peak signal-to-noise) in the literature. Our task is more challenging due to two factors. First, the pathology involved in source-modality

MR images significantly increases the difficulty of the synthesis task, since brain tumor varies in appearance, size and location. This is in contrast to the image synthesis for healthy subjects commonly seen in the literature. Second, synthesizing target-modality images only from source-modality has limits, since source-modality seems lack of some information observed in target-modality. For example, the diffuse changes in FLAIR images around tumor regions are not easily seem in T1 images, which may adversely affect the segmentation results. Therefore, in addition to our 3D cGAN, we propose to further polish the local details of our synthesized images through linearly combining the real target-modality images of the training set for approximation, and the combination weights are estimated from the target-modality-like images output by our 3D cGAN model. This approach is feasible because our target-modality-like images resemble the real ones so that their combination weights are highly correlated. Our method is both local and adaptive. "Local" means the combination weights vary with different locations in an image. "Adaptive" means the combination weights also change with different subjects.

Specifically, for a test subject that has only source-modality MR image, we partition its target-modality-like image from our 3D c-GAN into non-overlapping small patches ($16\times16\times16$) and approximate each patch $S^{te,gan}$ by the convex combination of the training patches $S_1^{tr,gan}, S_2^{tr,gan} \ldots, S_{N_{tr}}^{tr,gan}$ (N_{tr} denotes the number of training subjects) from the target-modality-like images at the same location. This is achieved by solving the following optimization problem:

$$\min_{w} \|\sum_{i=1}^{N_{tr}} w_i S_i^{tr,gan} - S^{te,gan}\|_2^2, s.t. \sum w_i = 1, w_i \geq 0. \tag{3.4}$$

The convex combination in Equation 3.4 assigns high weights to only a few very similar training patches, and near-zero weights to those dissimilar ones.

Due to the resemblance of our FLAIR-like images and the real ones, the above learned combination weights are further used to polish $S^{te,cc}$ by linearly combining the real FLAIR image patches $R_1^{tr}, R_2^{tr} \ldots, R_{N_{tr}}^{tr}$ at the same location in the training set:

$$S^{te,cc} = \sum_{i=1}^{N_{tr}} w_i R_i^{tr}. \tag{3.5}$$

Please note that the linear combination gives even better results than local non-linear mapping as shown in our experimental study. Although this convex combination imputes some artefacts that affect the appearance of the synthesized images, it proves to be an effective strategy to improve the segmentation, which is our ultimate target.

3.3.3 Brain Tumor Segmentation Model

To effectively segment brain tumor with the synthesized target-modality-like and source-modality MR images, we utilize an 11-layer, two-pathway 3D CNN segmentation model, DeepMedic [107], which achieves the state-of-the-art performance on brain tumor segmentation. The first pathway extracts small patches ($17\times17\times17$) with normal resolution. The second one processes context in low-resolution patches ($19\times19\times19$) down-sampled from the actual area of size $51\times51\times51$. Using this two-pathway structure can investigate different scales of input MR images to cater for both local and contextual image features.

In this chapter, the above segmentation model takes two channels of input: source-modality and target-modality. We train the model in two steps. In the first step, the source-modality and real target-modality images of the training samples are used for training in the usual way. Then, in the second step, this model is further fine-tuned with the source-modality and our synthesized target-modality images of the training samples. The fine-tuning is essential because for a given test sample, it is the synthesized target-modality image rather than the real unknown target-modality image that is used for segmentation. Please note that, the target-modality-like image of a training sample is generated by 3D cGAN and the local patch combination (section 3.3.2) using all the training samples excluding itself.

3.4 Experimental Result

3.4.1 Data and Experimental Setting

We evaluate our framework on the data set BRATS 2015 [14]. It consists of 274 subjects with the image size $240\times240\times155$ and four modalities: T1, T1C, T2 and FLAIR. Tumors are annotated as: 1) necrotic core, 2) edema, 3) non-enhancing and 4) enhancing core. We randomly select 230 subjects as training samples, and the rest as our test set. For each sample, two modalities, T1 and FLAIR, are used. The whole tumor and the tumor core part (classes 1,3,4) are segmented. We linearly scale the original intensity values in all images to $[-1,1]$ according to [107], without any additional contrast.

The learning rate of our 3D cGAN is fixed as 0.0002 in the first 100 training epochs and then linearly decays to zero in the next 100 epochs. Adam solver with batch-size six is applied, and λ is fixed as 300 when training 3D cGAN. Fine-tuning the segmentation model is executed with 15 epochs.

We evaluate our method from both the synthesis quality and the tumor segmentation performance. In this chapter, to evaluate synthesis quality, we use PSNR and normalized mean squared error (NMSE) to measure only the brain part and tumor part of images. To evaluate segmentation performance, we report dice scores (DSC) for both the whole tumor and the tumor core part.

3.4.2 Results and Discussion

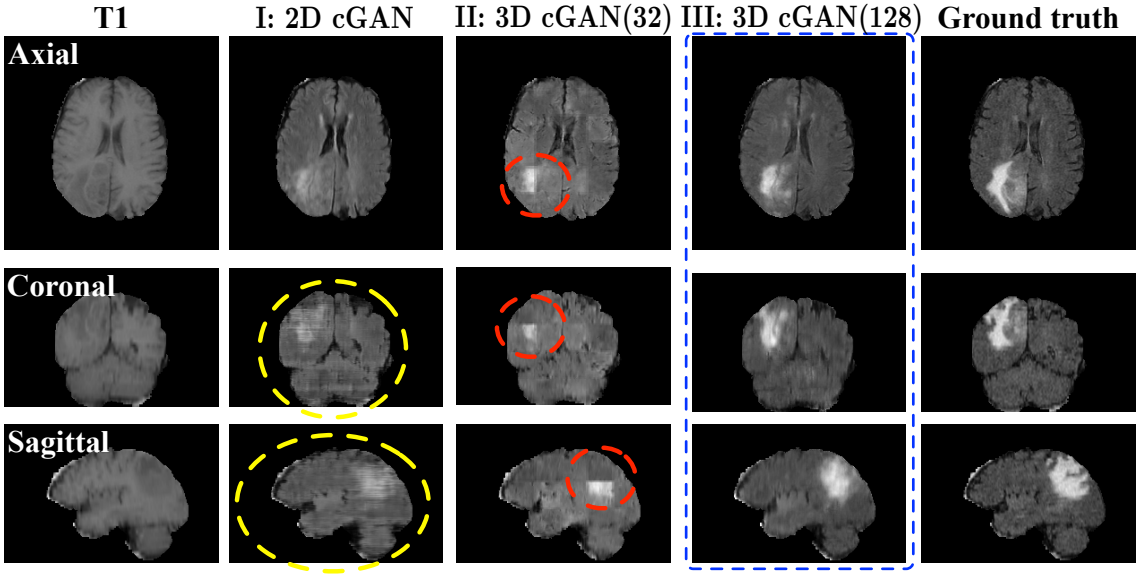


Figure 3.4: Visual comparison of the synthesized FLAIR images by 2D cGAN, 3D cGAN (32) and 3D cGAN (128). Discontinuity in the coronal and sagittal slices (in the yellow circles) is significant when using 2D cGAN. 3D cGAN (32) performs worse in the tumor region (in the red circles) when compared with 3D cGAN (128).

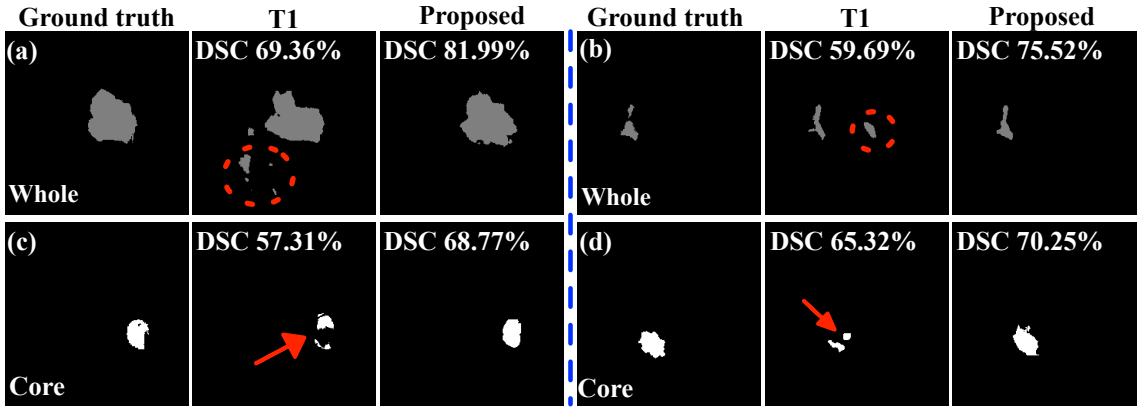


Figure 3.5: *Four patients of the segmented whole tumor and core parts by single T1 modality and our proposed method.*

To evaluate the effectiveness of our proposed method, we compare it with **I**: 2D cGAN on the whole zero-padded axial slices (256×256) [30], **II**: 3D cGAN on small patches (32×32×32), **III**: our proposed 3D cGAN on large patches (128×128×128), **IV**: local non-linear mapping (3D cGAN on 32×32×32-size patches) applied after the method **III**, and **V**: whole-image refinement by concatenating 3D cGANs like [123]. Note that the method **III** and **V** are the reduced variants of our proposed method. They are compared in this experiment to better demonstrate the effectiveness and advantage of the proposed method. Performance of the above-mentioned methods on the image synthesis and tumor segmentation is summarized in Table 3.1.

Table 3.1: Quantitative evaluation results of the synthesized brain and brain tumor segmentation. Numbers with underline indicate they are statistically significantly different from our proposed method (**III**+local adaptive fusion), according to a two-sided, paired t-test (solid line $p < 0.05$, dotted line $p < 0.1$)

Methods	**Synthesis quality** (PSNR/NMSE%)		**Segmentation performance** (DSC%)	
	brain	tumor	whole tumor	core
I: 2D cGAN	19.45/25.17	17.70/15.10	59.83	66.59
II: 3D cGAN (32)	19.53/25.12	18.09/13.05	52.06	50.02
III: **3D cGAN(128)**	20.45/25.08	19.13/12.68	66.35	72.09
IV: 3D cGAN (128)+3D cGAN(32)	19.94/24.99	18.73/13.45	66.61	72.14
V: 3D cGAN (128)+3D cGAN (128)	20.23/25.52	19.11/12.85	66.59	67.83
T1	N.A.		67.18	63.00
T1+real FLAIR (ideal scenario)	N.A.		*82.17*	*85.49*
III+local adaptive fusion)	**20.68/22.67**	**19.27/11.86**	**68.23**	**72.28**

From Table 3.1, it can be seen that the synthesized images by our method show the best quality with the highest PSNR and the lowest NMSE. This result is better than that of the 2D cGAN and all other 3D cGAN settings. Also, as shown, all the 3D cGAN based methods outperform the 2D cGAN in terms of synthesis quality, indicating the importance of considering 3D information during image synthesis. In addition, using the concatenated 3D cGANs (**V**) cannot further improve the synthesis results. Figure 3.4 shows an example of the visual comparison between the synthesized FLAIR-like images by **I**, **II** and our 3D cGAN (**III**). When looking into the coronal and sagittal slices, it is found that the discontinuity along these two directions is more significant in **I** than in **II** and **III**. Furthermore, it can be seen that using large patches as in our proposed 3D cGAN (**III**) can better synthesize the tumor parts than **II** that uses small patches.

The above observations are further supported by the segmentation performance in Table 3.1. Again, our method achieves the highest dice score of 68.23%, which is more than eight percentage points higher than that of the 2D cGAN. Our performance is also better than **III** that merely uses our 3D cGAN, indicating the essence of employing the proposed local adaptive fusion strategy. It is worthy noting that our method also beats **IV** that utilizes local non-linear fusion rather than our local linear one, in both synthesis and segmentation performance. And either non-linear or linear local fusion can further improve the results of 3D cGAN.

Table 3.1 also gives the dice scores of using the single modality of T1. As shown, jointly considering T1 and our synthesized FLAIR images can improve the tumor core part segmentation from 63% to 72.28% (ours wins in 33 subjects, and loses in 11 subjects), and the whole tumor segmentation from 67.18% to 68.23% (ours wins in 28 subjects, and loses in 16 subjects) compared with using T1 only. This suggests that our synthesized FLAIR-like images may carry complementary information about the soft tissue changes, which is helpful for T1-based brain tumor segmentation. Figure 3.5 displays four patients of the segmented whole tumors and core parts by T1-only method and our proposed one. As shown, when using T1 images only, some healthy brain tissues are segmented as parts of tumors (in the red circle) and some tumor core parts (indicated by the red arrow) are missing. The results are improved when our additional synthesized FLAIR images are jointly considered. In addition, in the ideal scenario, when the real FLAIR images are available with T1, the dice score could go up to 82.17%, indicating the large potential of our research direction.

In summary, both the visual and quantitative results demonstrate the capacity of our method in synthesizing FLAIR images from T1, and the benefits of using our synthesized FLAIR images to improve T1-based brain image segmentation, especially for tumor core parts.

3.5 Conclusion

Multi-modality MR images are crucial to achieve accurate brain tumor analysis. This chapter investigates how to synthesize high-quality target-modality MR images from the given source-modality. Through using the Unet-like generator, the proposed 3D cGAN model possesses the learning ability of extracting both local and global image feature representations in the global-level synthesis. With the 3D structures in its generator and discriminator, it can conduct the continues estimation along all the three dimensions of MR images. Also, the local adaptive fusion scheme further boosts the synthesized image quality. The proposed method outperforms 2D cGAN model for the cross-modality brain MR image synthesis task.

Chapter 4

Edge-aware GANs for Cross-modality MR Image Synthesis

For the cross-modality MR image synthesis task, the 3D cGAN model from Chapter 3 performs better than the 2D GAN that is usually applied on generic images. However, similar to other state-of-the-art GAN models, it only attempts to minimize the pixel-/voxel-wise intensity difference between the synthesized and the real images. For accurate brain image analysis, the visual information of brain structure in MR images is vital. Thus, this chapter proposes to preserve this information during the synthesis. Specifically, novel 3D edge-aware cGANs that are trained by two different adversarial learning strategies to additionally ensure the brain structure preservation, are studied for the global-level MR image synthesis.

4.1 Motivation

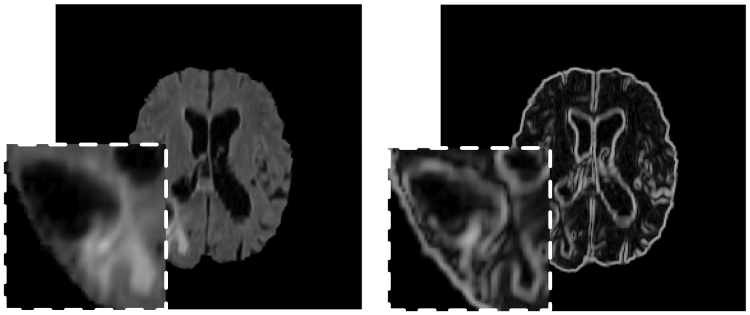
The CNNs based cGANs have well performed in generic image synthesis [30, 31]. Very recently, cGAN-based image synthesis models have also been applied to medical images, such as retinal images [125–127], CT images [55, 122, 128], PET images [58, 129], MR images [47, 60, 64, 66, 130–134], ultrasound images [135], and endoscopy images [136]. Most of these methods follow the work in [30] to simply minimize the pixel/voxel-wise difference between the synthesized and real images. This neglects the structural content in an image, such as the textures or shapes of objects, leading to less sharp synthesized images. The GAN model in [47] imposes additional constraints on the gradient similarity between real and synthesized images so that the sharpness of the synthesized images could be enhanced. Although this work is close to the proposed cGAN in this chapter, their fundamental differences will become clearer with the unfolding of the proposed model.

This thesis proposes edge-aware generative adversarial networks (Ea-GANs) to further

overcome the slice discontinuity and the less sharp synthesis problems in most of existing cGAN models for medical image synthesis. Our methods are 3D-based, and they extract both voxel-wise intensity and image structure information to facilitate synthesis. Since the information about local content and global context is crucial in the per-voxel prediction tasks, we design a Unet-like generator in the proposed Ea-GANs to exploit more effective feature representations about this information. More importantly, to capture the image structure details, we extract the edges that contain critical textural information for visual recognition [6, 137, 138], and integrate the edge maps with the adversarial learning in the cGAN model to boost synthesis quality. Specifically, two frameworks, a generator-induced Ea-GAN (gEa-GAN) and a more advanced discriminator-induced Ea-GAN (dEa-GAN), are proposed to learn Ea-GANs via different learning strategies. Note that, our method is significantly different from the gradient loss based method in [47]. The edge information provided by the Sobel operator is less sensitive to noise and favours nearer neighbours, compared with the direct use of gradient information. More importantly, the gradient information in [47] was only used in the objective function of the generator, but not involved in the adversarial learning like our dEa-GAN. The latter, however, is proven to be a very effective strategy to improve the synthesis quality. The Ea-GANs is validated on two datasets of MR images containing brain lesions and skulls, respectively. The effectiveness of the proposed methods is validated by comparing them with a set of state-of-the-art image synthesis methods [30, 41, 63]. Moreover, to show the generality of our edge-aware approach, we also test the 2D variant of the dEa-GAN on multiple generic 2D image synthesis tasks, which demonstrates consistent improvements over the methods in comparison.

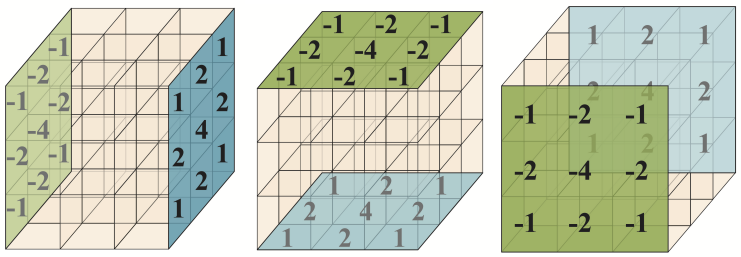
4.2 Proposed Ea-GANs

Most existing cGANs models, like Pix2pix [30], focus on the pixel-to-pixel/voxel-to-voxel image synthesis. They usually enforce the pixel/voxel-wise intensity similarity between the synthesized and real images. However, they ignore the structure of image content, such as the textural details in a MR image [139]. Since edges reflect the local intensity changes and show the boundaries between the different tissues in a MR image, maintaining edges can capture the textural structure of image content and help sharpen the synthesized MR images. Especially, when lesions are contained in MR images, the edge information helps differentiate the lesion and the normal tissues, and contributes to better depicting the contour of abnormal regions, e.g. gliomas tumors in brain MR images [140] (shown in the zoomed parts of Figure 4.1). To enforce edge preservation during MR image synthesis, we add an extra constraint based on the similarity of the edge maps from synthesized and real images. The edge maps are computed using the commonly used Sobel operator due to its simplicity, and its derivative can easily be computed for the



(a) FLAIR Image **(b)** Edge map of FLAIR Image

Figure 4.1: A brain FLAIR image (left), and the corresponding edge map (right) after the 3D Sobel edge detection. The contour of abnormal tissues can be depicted clearer by the edge map, which is shown as the zoomed regions.



(a) F_i **(b)** F_j **(c)** F_k

Figure 4.2: The three-dimensional Sobel operator includes three kernels as F_i, F_j, and F_k, respectively. The size of each kernel is $3 \times 3 \times 3$. Each empty cube without any number on its surface means the value of zero in the corresponding position of kernel. Similarly, the numbers in the blue or green cubes are the positive and negative values of three kernels.

backpropagation.

As shown in Figure 4.2, three Sobel filters, F_i, F_j, and F_k, are used to convolve an image A to generate three edge maps corresponding to the intensity gradients along i, j, and k directions, respectively. Then, these three edge maps are merged into a final edge map $S(A)$ by the following equation:

$$S(A) = \sqrt{(F_i * A)^2 + (F_j * A)^2 + (F_k * A)^2}, \tag{4.1}$$

where $*$ means the convolution operation.

Based on different strategies to utilise the edge maps, two frameworks, i.e., gEa-GAN and dEa-GAN, are proposed (as shown in Figure 4.3). Each of them consists of three modules, a generator G, a discriminator D, and a Sobel edge detector S. The details of these two frameworks are presented as follows.

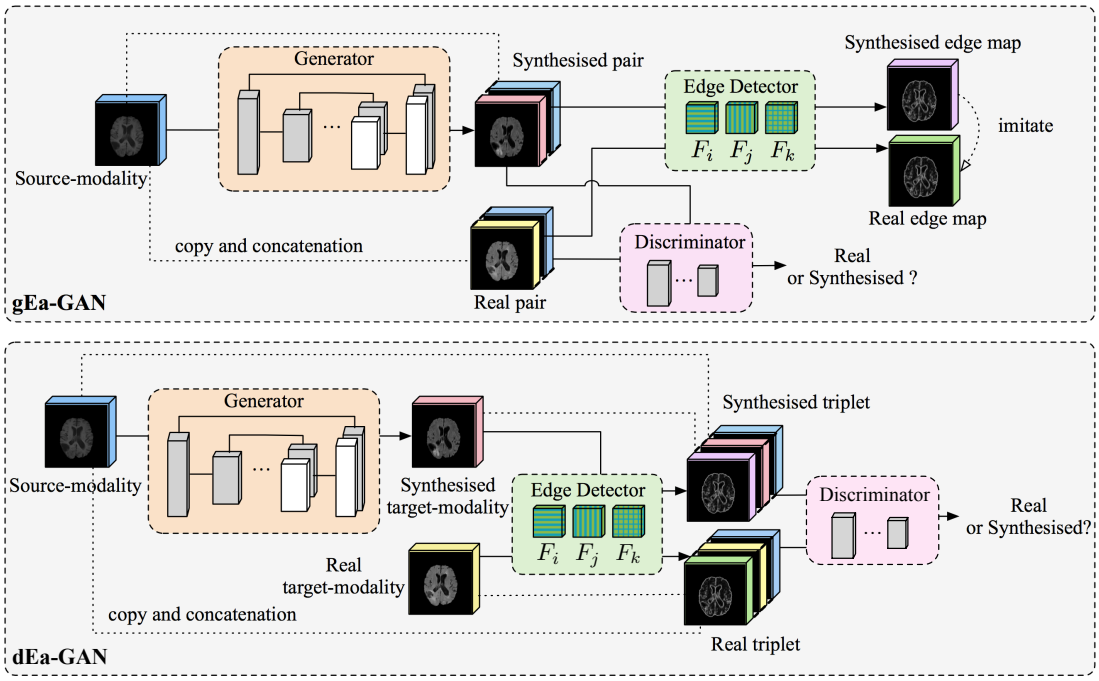


Figure 4.3: Frameworks of Ea-GANs. Both gEa-GAN and dEa-GAN include a generator G, a discriminator D, and a Sobel edge detector S. The generator G is trained to synthesize a realistic target-modality image with its edge map detected by the Sobel edge detector S, while the discriminator D is learned to distinguish between the synthesized and real pair/triplet for sharp synthesis. In the back-propagation step of training, the generator G of gEa-GAN is affected by the gradients from the dissimilarity between the synthetic and real edge maps, while both generator G and discriminator D of dEa-GAN are affected by the detected edge maps. The detailed architectures of generator and discriminator are given in Figures 3.2 and 3.3 of Chapter 3.

Generator-induced Ea-GAN

For cross-modality MR image synthesis task, a source-modality image $\mathbf{x} \sim p_{data}(\mathbf{x})$ and a target-modality image $\mathbf{y} \sim p_{data}(\mathbf{y})$ are scanned on the same subject with different contrasts. The generator G of the proposed gEa-GAN aims to synthesize target-modality-like images $G(\mathbf{x})$ that can fool its discriminator D by training with the adversarial loss. Also, the L1-norm penalties are applied through G to discourage the dissimilarity between the real and synthesized images, and between their edge maps extracted by the Sobel edge detector S. The constraint of edge map similarity is totally ignored in the cGAN, like Pix2pix, reviewed in Section 2.2. In this way, both of the voxel-wise intensity similarity and the edge similarity are enforced during the synthesis. Accordingly, the objective of its generator G is defined as follows:

$$\begin{aligned}\mathscr{L}^{G}_{gEa-GAN} = \mathbb{E}_{x\sim p_{data}(\mathbf{x})}[\log(1-D(\mathbf{x},G(\mathbf{x})))+ \\ \lambda_{l1}\mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{data}(\mathbf{x},\mathbf{y})}[\|\mathbf{y}-G(\mathbf{x})\|_1]+ \\ \lambda_{edge}\mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{data}(\mathbf{x},\mathbf{y})}[\|S(\mathbf{y})-S(G(\mathbf{x}))\|_1],\end{aligned} \tag{4.2}$$

where the hyper-parameters, λ_{l1} and λ_{edge}, are used to balance the three terms in Equation 4.2.

Following that in Pix2pix [30], the objective function of the discriminator D is defined as follows:

$$\begin{aligned}\mathscr{L}^{D}_{gEa-GAN} = -\mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{data}(\mathbf{x},\mathbf{y})}[\log D(x,y)] - \\ \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log(1-D(\mathbf{x},G(\mathbf{x})))].\end{aligned} \tag{4.3}$$

Finally, the gEa-GAN integrates its generator G and discriminator D together by training these two modules simultaneously with the following objective:

$$\mathscr{L}_{gEa-GAN} = \mathscr{L}^{G}_{gEa-GAN} + \mathscr{L}^{D}_{gEa-GAN}. \tag{4.4}$$

Discriminator-induced Ea-GAN

The gEa-GAN enforces the voxel-wise intensity similarity and the edge similarity by its generator for image synthesis. However, as the edge term only appears on the generator side, edge information is not perceived by the discriminator. Inspired by the mechanism of adversarial learning between the generator and discriminator, we further propose a dEa-GAN framework to incorporate the edge maps into this battle. Both generator and discriminator could benefit from the synthesized image and its edge map. Thus, the discriminator is also able to utilize the edge details to differentiate the real and synthesized images, and this in turn enforces the generator to produce the better edge details for synthesis.

Similar to the gEa-GAN, the generator G in the dEa-GAN model is trained using the adversarial loss, the voxel-wise intensity difference loss, and the edge difference loss for synthesis, according to the following objective:

$$\begin{aligned}\mathscr{L}^{G}_{dEa-GAN} = \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log(1-D(\mathbf{x},G(\mathbf{x}),S(G(\mathbf{x})))) + \\ \lambda_{l1}\mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{data}(\mathbf{x},\mathbf{y})}[\|\mathbf{y}-G(\mathbf{x})\|_1] + \\ \lambda_{edge}\mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{data}(\mathbf{x},\mathbf{y})}[\|S(\mathbf{y})-S(G(\mathbf{x}))\|_1].\end{aligned} \tag{4.5}$$

Compared with the gEa-GAN model, the edge map $S(G(\mathbf{x}))$ also implicitly appears in the first term of Equation 4.5 through the output of the discriminator D.

The objective of the discriminator D now becomes:

$$\begin{aligned}\mathscr{L}^{D}_{dEa-GAN} = -\mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{data}(\mathbf{x},\mathbf{y})}[\log D(\mathbf{x},\mathbf{y},S(\mathbf{y}))] - \\ \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log(1-D(\mathbf{x},G(\mathbf{x}),S(G(\mathbf{x}))))].\end{aligned} \tag{4.6}$$

As can be seen, the discriminator takes a triplet as its input by adding the edge map $S(G(\mathbf{x}))$ or $S(\mathbf{y})$. For a synthesized triplet composed of $\mathbf{x}$, $G(\mathbf{x})$, and $S(G(\mathbf{x}))$, the label is zero; for a real triplet composed of $\mathbf{x}$, $\mathbf{y}$, $S(\mathbf{y})$, the label is one. The discriminator tries to differentiate these two types of triplets.

Again, the final objective of this dEa-GAN model is:

$$\mathscr{L}_{dEa-GAN} = \mathscr{L}^{G}_{dEa-GAN} + \mathscr{L}^{D}_{dEa-GAN}. \tag{4.7}$$

4.2.1 Detailed Architectures

The proposed two Ea-GANs are three-dimensional to capture the local and global contextual information along all three directions. Both of them consist of three modules, the generator G, the discriminator D, and the edge detector S. Among these three modules, the edge detector S is the Sobel operator, and the generator and discriminator are built upon CNN architectures to extract deep features from images. Due to limited GPU memory and the required training batch-size, we utilize large overlapped patches ($128 \times 128 \times 128$) rather than a whole image to train our Ea-GAN models, which could provide a sufficient number of samples to train a good model. Our generator and discriminator in Ea-GANs share the same architectures with those in 3D cGAN from Chapter 3 (Figures 3.2 and 3.3), except the first layers of their discriminators. For the proposed gEa-GAN model, the input of its discriminator is a pair of images, so the discriminator takes in two channels of 3D large patches. Meanwhile, the dEa-GAN model processes triplets with three channels of 3D large patches. Thus, the designed discriminators vary in the two Ea-GAN models, the gEa-GAN and the dEa-GAN, by their first layers to involve the different numbers of input channels, which are two for gEa-GAN and three for dEa-GAN.

4.2.2 Implementation

When training a GANs model, a common issue is that it can become unstable or even experience a mode collapse easily [141]. For example, the discriminator tends to be more powerful than the generator, which is reflected by different decreasing speeds of their loss functions. In this case, the whole model is unstable, and cannot synthesize high-quality images. Many techniques have been discussed in [141] to improve the stability of training the GAN models. In our work, we consider two strategies. First, the labels used by the discriminator are smoothed to raise the difficulty of differentiation, and further reduce the vulnerability of adversarial learning. For a synthesized pair/triplet, the target label of the discriminator is set as a random number between 0 and 0.3, while for real pair/triplet, the target label is set as a random value between 0.7 and 1.2, inspired by [142]. In this way, the task of discriminator becomes more challenging to match the difficulty of the task of generator, so that the adversarial training becomes balanced.

The second strategy is used to better utilize the edge information in MR images. At the initial stage of training, the quality of the extracted edge maps, which highly relies on the generated images, is not good enough to effectively guide the synthesis. To mitigate this issue, the value of the hyper-parameter λ_{edge} is initially set to be small and then gradually increased to adjust the importance of edge information. Specifically, we linearly increase λ_{edge} from zero to 100 in the first 20 epochs, and then fix it at 100 in the following epochs. In this way, the Ea-GANs can effectively utilize the edge information to synthesize sharp and realistic target-modality-like images.

4.3 Experimental Results

4.3.1 Dataset and Experimental Setting

We use two datasets, i.e. the brain tumor contained BRATS2015 [14] and the non-skull stripped IXI [15], to evaluate two proposed Ea-GANs.

The BRATS2015 dataset consists of 274 subjects with four modalities of co-registered MR images: T1-weighted (T1), T1-weighted and contrast-enhanced (T1c), T2-weighted (T2) and FLAIR, with the image size $240\times240\times155$ (voxels). In this chapter, we use T1 as the source-modality since it is the most commonly used modality for structural imaging, and test two synthesis tasks with FLAIR and T2 as the target-modality, respectively. Note that, different from Chapter 3, this chapter applies five-fold cross-validation to effectively evaluating different methods on the entire dataset. For each cross-validation split, we divide the dataset into a training set (consisting of 4/5 samples) and a test set (consisting of 1/5 samples). The original intensity values of all used images are linearly scaled to $[-1,1]$ without any additional contrast change before processed by Ea-GANs. For each image, eight large patches (size: $128\times128\times128$) are extracted, and the overlapped regions are averaged to form the final estimation. Note that, in order to increase the number of training samples, we use large patches rather than whole images for training, which is essentially different from voxel-wise regression used in the traditional small patch-based synthesis methods.

The IXI dataset includes 578 subjects of the non-skull stripped brain MR images from five modalities, i.e., T1, T2, proton density (PD), MRA, and diffusion tensor imaging (DTI), with the image size $256\times256\times N$ (N is different for each subject). We synthesize T2 images from PD images following [63]. The dataset is utilized by a five-fold cross-validation, so the training set and test set consist of samples from 4/5 and 1/5 subjects, respectively, for each cross-validation split. We also linearly scale the original intensity values into $[-1,1]$ without any additional contrast change in the pre-processing. For each 3D image, non-overlapped large patches (size: $128\times128\times128$) are extracted along the trans-coronal and trans-sagittal directions. Along the trans-axial direction, patches are

padded with -1 if $N < 128$.

For all synthesis tasks, we conduct 150 epochs to train the models. In the first 100 epochs, the leaning rate is fixed as 0.0002, and then it linearly decays to zero in the next 50 epochs. Adam solver with a batch-size of six is applied to minimize the objectives. During training, λ_{l1} is fixed as 300, while λ_{edge} linearly increases from zero to 100 in the first 20 epochs, and then stays at 100 in the next 130 epochs. Before the evaluation, the intensity values of all the synthesized and real images are added by one, and then divided by two. Thus, the intensity values of all images are between zero and one in measure metrics.

4.3.2 Methods in Comparison

The proposed Ea-GANs are compared with two state-of-the-art cross-modality MRI synthesis methods, Replica [41] and Multimodal [63], and a generic-image-synthesis cGAN-based model, Pix2pix [30].

1. Replica [41] uses the handcrafted multi-resolution 3D patch features to train random forests for synthesis.

2. Multimodal [63] is a 2D CNN-based model to synthesize the MR image slice by slice with the constraint of pixel-wise intensity difference.

3. Pix2pix [30] is a 2D cGAN model, which synthesizes whole 2D images by focusing on maintaining the pixel-wise intensity similarity.

We directly run these three models using their publicly released codes and follow the original papers for both image pre-processing and model setting. The two 2D models, i.e., Multimodal [63] and Pix2pix [30], are trained using axial slices. Then the synthesized axial slices of each subject are concatenated to form a 3D volume.

Moreover, to facilitate the comparison, the proposed 3D cGAN model from Chapter 3 has the same architecture and parameter setting as the two proposed models, i.e., gEa-GAN and dEa-GAN. All of the 3D models work on large 3D patches to increase the number of training samples.

In addition, to verify the advantages of using edge maps over directly using image gradients, we also build a cGAN model that follows the network architecture of gEa-GAN and dEa-GAN but use the image gradient difference loss in [47] instead of the edge similarity loss. This model is denoted as a gradient cGAN. To balance each term in the objective function, the added image gradient difference loss is normalized by the number of output voxels and multiplied by 3000.

4.3.3 Results on BRATS2015

The synthesis results on the BRATS2015 dataset are reported, including those from Replica [41], Multimodal [63], Pix2pix [30], 3D cGAN, gradient cGAN, and the two proposed methods of gEa-GAN and dEa-GAN. We show their synthesis performances evaluated on whole images (including a brain and a background) in Table 4.1 and on tumor regions in Table 4.2, respectively. To test if the proposed dEa-GAN is statistically significantly better than a compared method, paired t-test is conducted, following [143–145]. When the improvement of dEa-GAN over the method is statistically significant, the result of that compared method will be underlined in the tables. The visual examples for the two tasks are correspondingly presented in Figures 4.4 and 4.5.

Comparison between 2D and 3D cGANs

In this chapter, we further compare 3D cGAN (from Chapter 3) with 2D cGAN for two synthesis tasks on the entire BRATS2015 dataset. As shown in Table 4.1, 3D cGAN which is a 3D variant of Pix2pix [30], significantly outperforms Pix2pix [30] with better synthesis results. Specifically, this 3D model improves the quality of T1-to-FLAIR synthesis and T1-to-T2 synthesis by (1) 1.8dB PSNR, 0.025 NMSE, and 0.018 SSIM (T1 to FLAIR), and (2) 1.22dB PSNR, 0.015 NMSE, and 0.011 SSIM (T1 to T2), respectively, compared with the 2D model of Pix2pix [30]. This demonstrates the importance of considering 3D contextual information during the synthesis. When looking into the coronal and sagittal slices in Figure 4.4 and Figure 4.5, it is found that the discontinuity along these two directions is more salient in Pix2pix [30] than in 3D cGAN (especially the regions between two red arrows in the zoomed parts).

Comparison between 3D cGAN and the Proposed Ea-GANs

As a model only focusing on the voxel-wise intensity similarity, 3D cGAN produces lower-quality images than both of the proposed Ea-GANs which jointly consider the intensity similarity and edge similarity during training. Concretely, for the synthesis task of FLAIR, PSNR and SSIM increase from 29.26dB (3D cGAN) to 30.11dB (dEa-GAN) and from 0.958 (3D cGAN) to 0.963 (dEa-GAN) respectively, and NMSE decreases from 0.119 (3D cGAN) to 0.105 (dEa-GAN). For T2 synthesis results, dEa-GAN improves PSNR, NMSE, and SSIM by 0.61db, 0.007, and 0.003 from 3D cGAN, respectively. These improvements consistently demonstrate the necessity of preserving edge details in image synthesis. Meanwhile, from all the three views in Figure 4.4 and Figure 4.5, the proposed two Ea-GANs synthesize sharper edges than the 3D cGAN (indicated by two blue arrows in the zoomed parts).

Table 4.1: Quantitative evaluation results of the synthesized FLAIR-like and T2-like images from T1 on the BRATS2015 dataset (mean±standard deviation). The paired t-test is conducted between dEa-GAN and a compared method at the significance level of 0.05. When the improvement of dEa-GAN over the method is statistically significant, the result of that compared method will be underlined.

Methods	**T1 to FLAIR**			**T1 to T2**		
	PSNR	NMSE	SSIM	PSNR	NMSE	SSIM
Replica [41]	27.17±2.60	0.171±0.267	0.939±0.013	26.92±2.36	0.158±0.324	0.946±0.015
Multimodal [63]	27.26±2.82	0.184±0.284	0.950±0.014	27.31±2.39	0.140±0.229	0.951±0.016
Pix2pix [30]	27.46±2.55	0.144±0.189	0.940±0.015	28.12±2.45	0.110±0.220	0.953±0.014
3D cGAN (from Chapter 3)	29.26±3.21	0.119±0.205	0.958±0.016	29.34±3.23	0.095±0.199	0.964±0.017
gradient cGAN (ablation study)	29.38±3.25	0.116±0.204	0.960±0.017	29.43±3.28	0.097±0.210	0.966±0.017
Proposed gEa-GAN	29.55±3.24	0.115±0.199	0.960±0.017	29.58±3.29	0.093±0.218	0.966±0.018
Proposed dEa-GAN	**30.11±3.22**	**0.105±0.174**	**0.963±0.016**	**29.98±3.37**	**0.088±0.223**	**0.967±0.016**

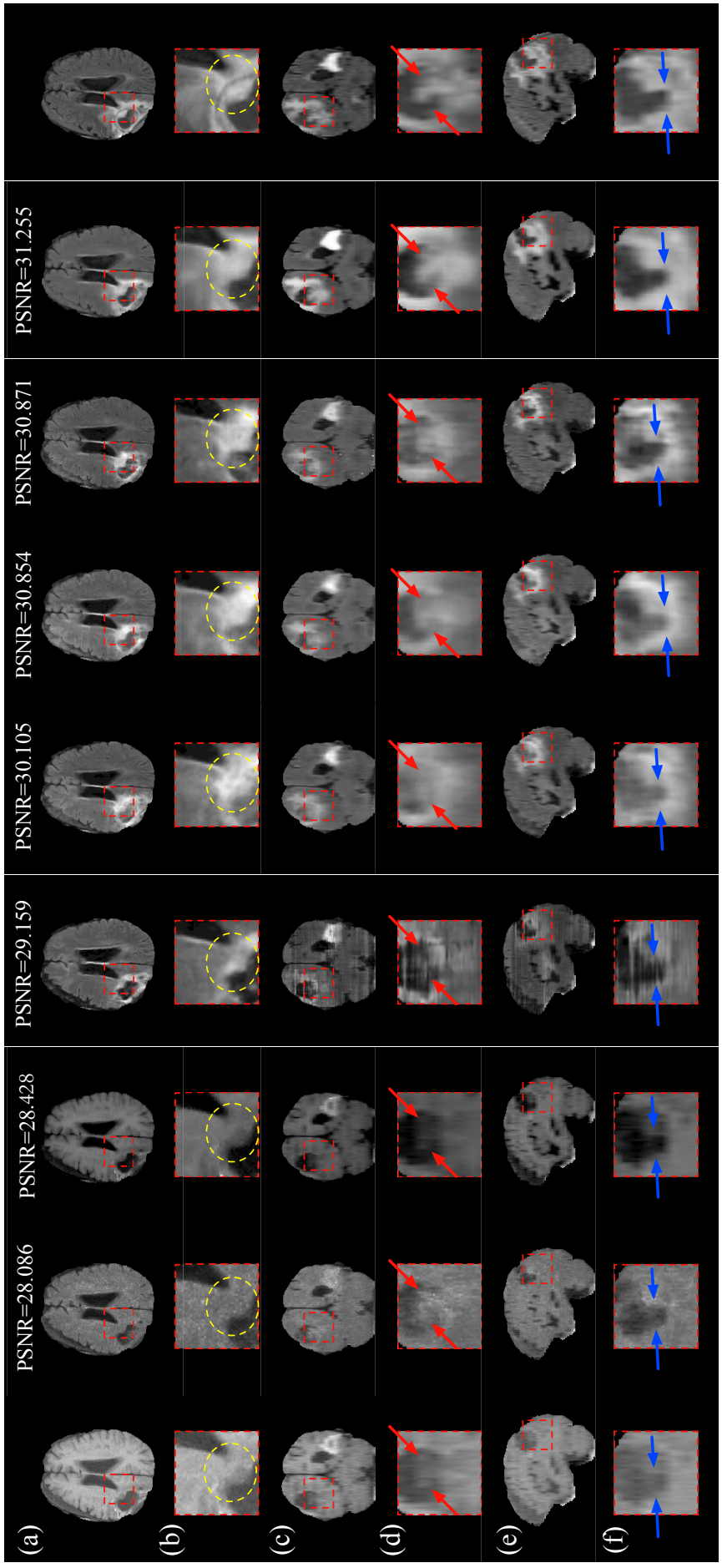


Figure 4.4: Comparison between the two proposed Ea-GANs and other state-of-the-art methods (T1 to FLAIR on the BRATS2015 dataset): (a) axial slices, (b) zoomed parts of axial slices, (c) coronal slices, (d) zoomed parts of coronal slices, and (e) sagittal slices, (f) zoomed parts of sagittal slices.

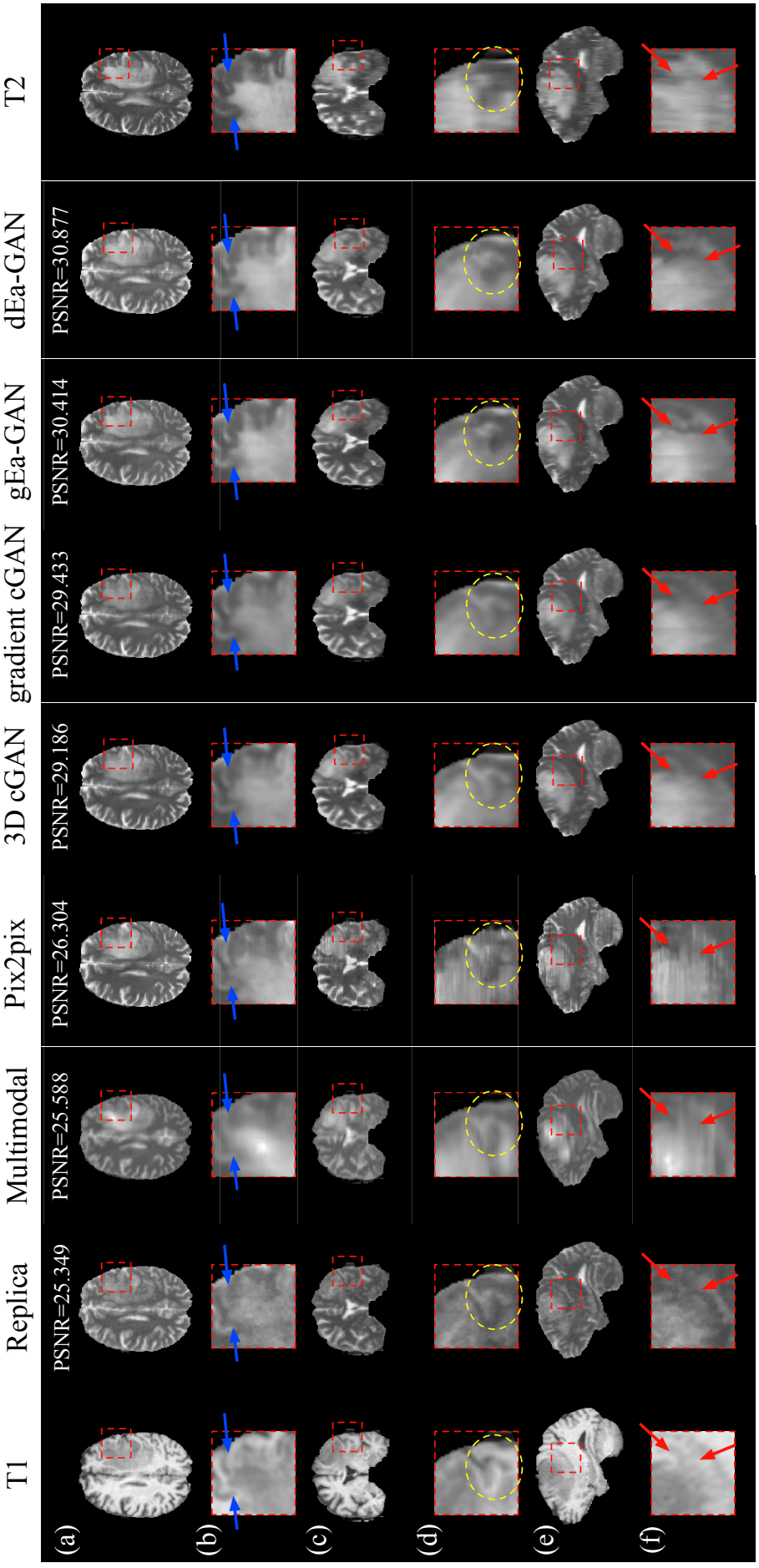


Figure 4.5: Comparison between the two proposed Ea-GANs and other state-of-the-art methods (T1 to T2 on the BRATS2015 dataset): (a) axial slices, (b) zoomed parts of axial slices, (c) coronal slices, (d) zoomed parts of coronal slices, and (e) sagittal slices, (f) zoomed parts of sagittal slices.

Comparison between Gradient cGAN and gEa-GAN

The gradient cGAN, which directly extracts the image gradient difference for training, performs worse than the proposed gEa-GAN, as shown in Table 4.1. These two methods only differ in whether gradient similarity or edge similarity is used. This indicates the superiority of using the Sobel edge similarity loss over directly using the image gradient loss. The performance is further improved when the edge similarity is also adversarially learned in dEa-GAN.

Comparison between the Two Proposed Ea-GANs

When comparing between the two proposed Ea-GANs, the dEa-GAN significantly improves the averaged PSNR and SSIM values by approximately 0.6dB and 0.003 respectively and lowers the NMSE value by about 0.01 from the gEa-GAN for the FLAIR synthesis task. Similarly, the dEa-GAN improves PSNR, NMSE, and SSIM by 0.4dB, 0.005, and 0.01, respectively, for the T2 synthesis task. Those validate that integrating the edge information into both generator and discriminator can significantly enhance the learning of edge similarity, and further improves the whole image synthesis performance.

Comparison between the State-of-the-art Models and the Proposed Ea-GANs

When comparing our results with the state-of-the-art methods in literature, Replica [41] obtains the worst PSNR and SSIM evaluation results, which indicates that the small-patch-based method with the handcrafted features may not be able to capture the image contextual information for MR synthesis. For the 2D models, Pix2pix [30] can get slightly better quantitative results than Multimodal [63]. However, when comparing them with the proposed Ea-GANs, our methods outperform these two methods in terms of all the three measures. As shown in the yellow circles of Figure 4.4 and Figure 4.5, our methods produce the FLAIR-like and T2-like images with the more local details along all the three directions: axial, sagittal, and coronal. It is worth noting that, the gradient cGAN can be regarded as the best performing one in the compared methods, in terms of either the closest mean or the smallest average difference to our proposed dEa-GAN. Still, our dEa-GAN performs statistically significantly better than it as analyzed above.

Comparison on Tumor Regions

The above seven methods are also compared on the lesion-contained regions in Table 4.2. The Ea-GANs obtain the best values of PSNR, NMSE, and SSIM over all the methods in comparison. This is consistent with our observations on whole images. Also, the dEa-GAN shows statistically significant improvements over the best-performing one, i.e., the gradient cGAN, among the compared methods. These show the capacity of the proposed

Table 4.2: Quantitative evaluation results of the synthesized FLAIR-like and T2-like tumor parts from T1 on the BRATS2015 dataset (mean±standard deviation). The paired t-test is conducted between dEa-GAN and a compared method at the significance level of 0.05. When the improvement of dEa-GAN over the method is statistically significant, the result of that compared method will be underlined.

Methods	**T1 to FLAIR**			**T1 to T2**		
	PSNR	NMSE	SSIM	PSNR	NMSE	SSIM
Replica [41]	13.34±3.41	0.137±0.068	0.601±0.083	14.93±3.17	0.123±0.084	0.650±0.139
Multimodal [63]	13.82±3.66	0.131±0.076	0.638±0.096	15.50±3.75	0.109±0.117	0.689±0.138
Pix2pix [30]	14.48±3.12	0.127±0.093	0.618±0.084	16.03±3.10	0.099±0.084	0.703±0.095
3D cGAN (from Chapter 3)	15.95±3.52	0.098±0.094	0.681±0.090	16.79±3.56	0.089±0.093	0.725±0.099
gradient cGAN (ablation study)	15.67±3.63	0.104±0.123	0.682±0.090	16.87±3.40	0.085±0.089	0.752±0.098
Proposed gEa-GAN	16.37±3.49	0.090±0.101	0.697±0.092	17.23±3.50	0.083±0.099	0.752±0.100
Proposed dEa-GAN	**16.90±3.59**	**0.084±0.099**	**0.705±0.093**	**18.02±3.55**	**0.079±0.098**	**0.766±0.098**

Ea-GANs on preserving the critical pathological information in the synthesized images, since such information could be correlated to edges.

4.3.4 Results on IXI Dataset

As can be seen in Table 4.3, the proposed Ea-GANs outperform the other five methods in comparison according to all the three measures. The proposed dEA-GAN model demonstrates considerable improvements, with NMSE dropping from 0.087 (Replica) to 0.031 (dEa-GAN), SSIM rising from 0.947 (Replica) to 0.977 (dEa-GAN), and PSNR rising from 27.99dB (Replica) to 33.25dB (dEa-GAN), respectively. The second best is the proposed gEa-GAN model. These results validate that the two proposed Ea-GANs can also synthesize the non-skull stripped MR images with higher quality. Example images are shown in Figure 4.6. Although all the methods produce the high-quality synthesized T2 images, the visual results generated by the two proposed Ea-GANs show the sharper edges (indicated by the two red arrows in the zoomed parts), which is consistent with the observation from quantitative evaluation.

4.3.5 Results on the Synthesized Image Edge Maps

To directly show the edge preserving performance of the proposed Ea-GANs, three kinds of edge maps, i.e. Sobel, Prewitt, and Canny binary edge maps, extracted from the synthesized and real images for the three synthesis tasks are compared via PSNR, NMSE, and SSIM in Tables 4.4, 4.5, 4.6, and 4.3. As shown, the proposed Ea-GANs produce the edge maps that are closest to the ground-truth. These results directly verify the effectiveness of maintaining edge similarity by the two proposed Ea-GANs.

4.3.6 Ablation Study

To further investigate the details of the proposed Ea-GANs, the ablation study is also designed and conducted from two aspects. The first one is to study the effect of hyper-parameters in the loss functions during training. The second one is to explore the number of epochs until the convergence. Considering the high computational costs for 3D models, the following ablation study results are evaluated on randomly selected 20%subjects from the corresponding datasets, and the rest 80% subjects are used for training.

Table 4.3: Quantitative evaluation results of the synthesized T2-like images and their edge maps from PD on the IXI dataset (mean±standard deviation). The paired t-test is conducted between dEa-GAN and a compared method at the significance level of 0.05. When the improvement of dEa-GAN over the method is statistically significant, the result of that compared method will be underlined.

Methods	**Whole image**			**PSNR of edge maps**		
	PSNR	NMSE	SSIM	Sobel	Prewitt	Canny
Replica [41]	27.99±1.65	0.087±0.034	0.947±0.013	29.43±1.77	29.35±1.76	13.33±0.57
Multimodal [63]	29.77±2.46	0.078±0.080	0.953±0.019	30.87±1.96	30.76±1.97	14.25±0.72
Pix2pix [30]	30.80±1.90	0.054±0.031	0.964±0.012	31.62±1.49	31.59±1.45	14.48±0.72
3D cGAN (from Chapter 3)	32.10±2.02	0.038±0.039	0.973±0.011	32.61±1.50	32.56±1.48	14.58±0.77
gradient cGAN (ablation study)	32.51±2.18	0.036±0.036	0.974±0.011	32.91±1.69	32.86±1.67	14.59±0.77
Proposed gEa-GAN	32.81±2.11	0.035±0.036	0.975±0.010	33.10±1.60	33.06±1.59	14.79±0.78
Proposed dEa-GAN	**33.25±2.20**	**0.031±0.032**	**0.977±0.011**	**33.59±1.69**	**33.55±1.67**	**15.20±0.78**

Methods	**NMSE of edge maps**			**SSIM of edge maps**		
	Sobel	Prewitt	Canny	Sobel	Prewitt	Canny
Replica [41]	0.154±0.039	0.153±0.039	0.682±0.080	0.947±0.013	0.949±0.012	0.860±0.018
Multimodal [63]	0.124±0.072	0.124±0.074	0.554±0.086	0.956±0.014	0.958±0.014	0.884±0.020
Pix2pix [30]	0.097±0.041	0.096±0.041	0.538±0.087	0.963±0.009	0.964±0.009	0.886±0.020
3D cGAN (from Chapter 3)	0.079±0.044	0.078±0.045	0.492±0.082	0.968±0.009	0.969±0.009	0.897±0.018
gradient cGAN (ablation study)	0.076±0.037	0.075±0.037	0.463±0.074	0.970±0.010	0.971±0.009	0.903±0.017
Proposed gEa-GAN	0.067±0.039	0.066±0.039	0.454±0.074	0.973±0.009	0.974±0.008	0.904±0.017
Proposed dEa-GAN	**0.064±0.034**	**0.063±0.034**	**0.445±0.073**	**0.975±0.009**	**0.976±0.008**	**0.906±0.017**

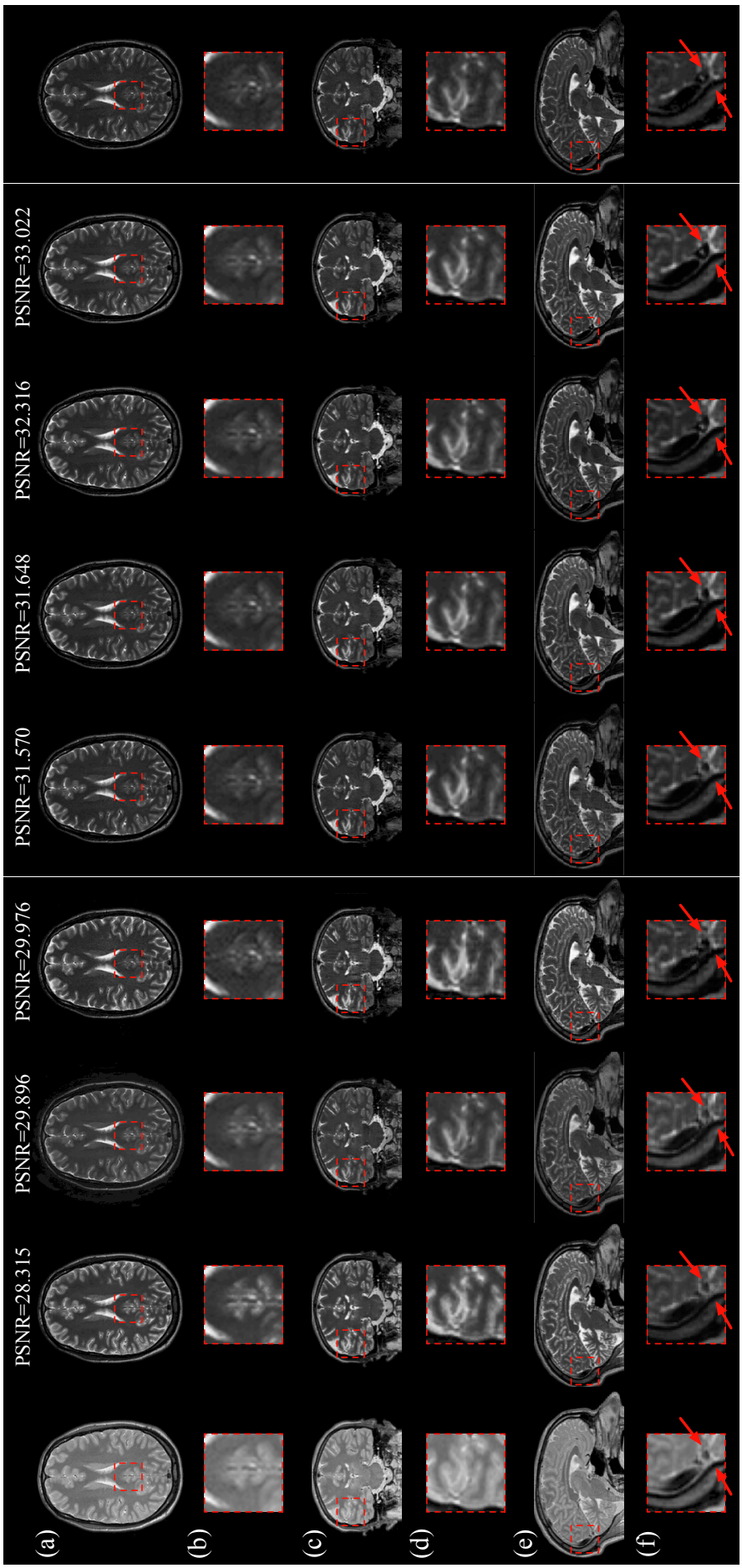


Figure 4.6: Comparison between the two proposed Ea-GANs and other state-of-the-art methods (PD to T2 on the IXI dataset): (a) axial slices, (b) zoomed parts of axial slices, (c) coronal slices, (d) zoomed parts of coronal slices, and (e) sagittal slices, (f) zoomed parts of sagittal slices.

Table 4.4: PSNR evaluation results of the synthesized edge maps on the BRATS2015 dataset (mean±standard deviation). The paired t-test is conducted between dEa-GAN and a compared method at the significance level of 0.05. When the improvement of dEa-GAN over the method is statistically significant, the result of that compared method will be underlined.

Methods	**T1 to FLAIR**			**T1 to T2**		
	Sobel	Prewitt	Canny	Sobel	Prewitt	Canny
Replica [41]	31.08±1.85	30.91±1.81	15.40±1.03	30.39±1.59	30.21±1.55	15.20±0.78
Multimodal [63]	31.40±1.95	31.27±1.93	15.61±1.11	30.16±1.72	29.96±1.73	15.18±0.80
Pix2pix [30]	31.64±1.63	31.46±1.60	15.73±0.84	31.51±1.46	31.34±1.47	15.22±0.76
3D cGAN (from Chapter 3)	32.87±1.95	32.73±1.95	16.48±0.95	32.53±1.84	32.37±1.87	15.70±0.90
gradient cGAN (ablation study)	33.06±2.07	32.91±2.05	16.64±0.98	32.63±1.97	32.67±2.00	15.77±0.92
Proposed gEa-GAN	33.10±2.04	32.96±2.03	16.64±0.97	32.87±1.98	32.70±2.01	15.78±0.92
Proposed dEa-GAN	**33.25±2.08**	**33.11±2.08**	**16.72±0.99**	**32.88±1.98**	**32.71±2.01**	**16.16±0.93**

Table 4.5: NMSE evaluation results of the synthesized edge maps on the BRATS2015 dataset (mean±standard deviation). The paired t-test is conducted between dEa-GAN and a compared method at the significance level of 0.05. When the improvement of dEa-GAN over the method is statistically significant, the result of that compared method will be underlined.

Methods	**T1 to FLAIR**			**T1 to T2**		
	Sobel	Prewitt	Canny	Sobel	Prewitt	Canny
Replica [41]	0.381±0.180	0.370±0.180	1.013±0.180	0.452±0.334	0.448±0.335	1.306±0.342
Multimodal [63]	0.372±0.230	0.367±0.232	1.019±0.117	0.445±0.428	0.444±0.436	1.111±0.181
Pix2pix [30]	0.369±0.225	0.366±0.228	1.021±0.250	0.290±0.185	0.284±0.183	1.099±0.177
3D cGAN (from Chapter 3)	0.292±0.229	0.287±0.231	0.971±0.253	0.238±0.172	0.233±0.171	0.992±0.197
gradient cGAN (ablation study)	0.279±0.217	0.275±0.218	0.937±0.245	0.226±0.170	0.222±0.169	0.976±0.211
Proposed gEa-GAN	0.277±0.237	0.274±0.238	0.939±0.252	0.225±0.177	0.221±0.176	0.969±0.206
Proposed dEa-GAN	**0.269±0.225**	**0.266±0.228**	**0.937±0.256**	**0.224±0.173**	**0.220±0.172**	**0.954±0.200**

Table 4.6: SSIM evaluation results of the synthesized edge maps on the BRATS2015 dataset (mean±standard deviation). The paired t-test is conducted between dEa-GAN and a compared method at the significance level of 0.05. When the improvement of dEa-GAN over the method is statistically significant, the result of that compared method will be underlined.

Methods	**T1 to FLAIR**			**T1 to T2**		
	Sobel	Prewitt	Canny	Sobel	Prewitt	Canny
Replica [41]	0.938±0.008	0.939±0.008	0.897±0.016	0.954±0.008	0.955±0.008	0.900±0.015
Multimodal [63]	0.947±0.012	0.949±0.014	0.905±0.016	0.955±0.014	0.956±0.013	0.900±0.016
Pix2pix [30]	0.959±0.009	0.960±0.008	0.909±0.015	0.963±0.009	0.964±0.009	0.905±0.015
3D cGAN (from Chapter 3)	0.965±0.010	0.966±0.010	0.923±0.015	0.948±0.010	0.969±0.010	0.913±0.017
gradient cGAN (ablation study)	0.965±0.010	0.966±0.010	0.925±0.015	0.970±0.011	0.971±0.010	0.915±0.017
Proposed gEa-GAN	0.968±0.010	0.969±0.010	0.924±0.015	0.970±0.010	0.971±0.010	0.915±0.017
Proposed dEa-GAN	**0.968±0.010**	**0.970±0.010**	**0.926±0.016**	**0.971±0.010**	**0.972±0.010**	**0.915±0.017**

Study about Hyper-parameters

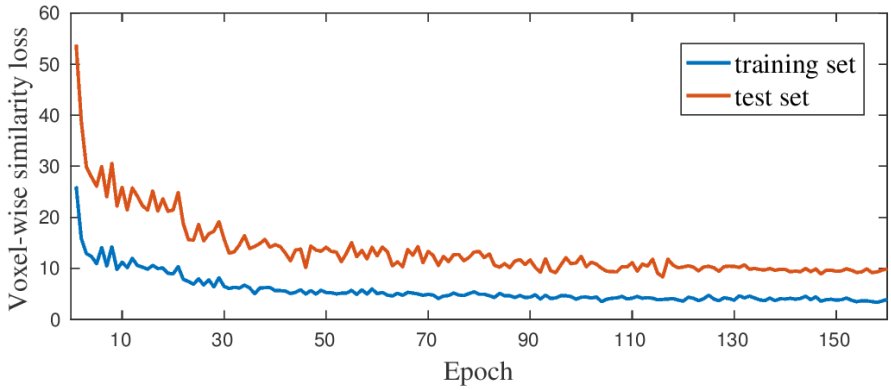
Table 4.7: Ablation study of λ_{l1} in 3D cGAN on BRATS2015 dataset (T1 to FLAIR)

λ_{l1}	PSNR	NMSE	SSIM
100	29.40	0.122	0.956
200	29.31	0.119	0.955
300	29.60	0.121	0.959
400	29.46	0.122	0.957

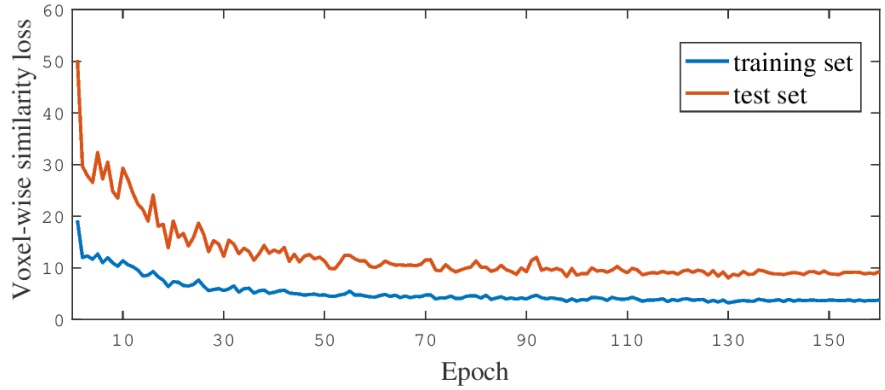
We conduct the ablation study of 3D cGAN with different λ_{l1}s on the BRATS2015 dataset (T1 to FLAIR) to show the influence of λ_{l1}. Compared to the adversarial loss, the values of the L1-norm based similarity loss and the edge cost are very small. Also, considering the computational time for training 3D CNNs models, we test 3D cGAN models with λ_{l1} of 100, 200, 300, and 400 as the typical cases. As shown in Table 4.7, there are no significant differences in synthesis results by using the different λ_{l1}s. Thus, we choose the relatively better one, i.e., 300, for λ_{l1} in all the synthesis tasks. At the same time, to balance the voxel-wise similarity loss and the edge similarity loss, we then set λ_{edge} as 100 during the fixed λ_{edge} training stage. Please note that these parameters are uniformly applied to all the datasets, and they show the good generalization.

Study about Training Epochs

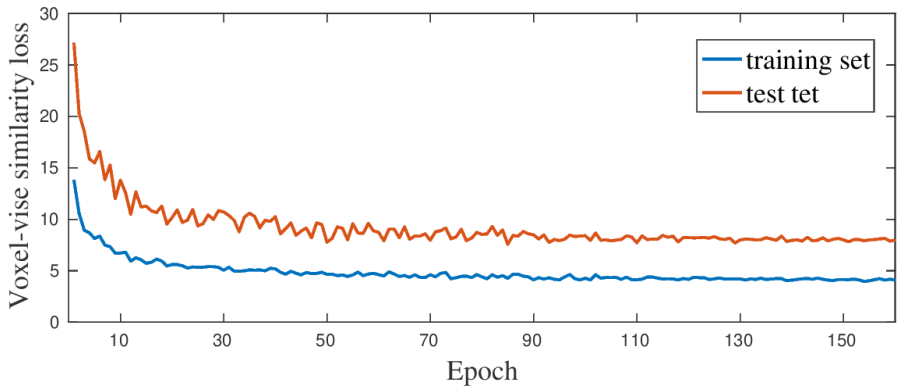
To study the effect of training epochs in training the proposed model, its training error, i.e., voxel-wise similarity loss, is employed, as presented in Figure 4.7. Specifically, the relationship between the training error and the epoch number is plotted for the BRATS2015 dataset for both T1-to-FLAIR (in Figure 4.7 (a)) and T1-to-T2 (in Figure 4.7 (b)) tasks, and also for the IXI dataset (Figure 4.7 (c)). As can be seen in Figure 4.7 (a), with the increase of training epochs, the losses for both the training and the test samples are converging to some constant numbers. After 130 epochs, both the training and the test losses become stable at the small values, so the trained model does not under-fit at this point. Meanwhile, the test loss does not increase with the decrease of the training loss, which means the problem of over-fitting does not happen. For all the other two tasks in Figure 4.7 (b) (c), the training errors also converge in about 150 epochs. By applying this epoch number setting (determined solely based on training errors), the obtained test errors also converge well, demonstrating that this is a reasonable choice. It is worth mentioning that the same epoch number is uniformly applied on both the BRATS2015 dataset (including two different tasks) and the IXI dataset in our experiments, rather than tweaking this epoch number for each dataset individually. In IXI dataset, the MR images are PD-contrasted, not skull-stripped and do not contain tumors, and therefore has different



(a) BRATS2015 (T1-to-FLAIR)



(b) BRATS2015 (T1-to-T2)



(c) IXI (PD-to-T2)

Figure 4.7: Ablation study: plot of the objective values versus the epoch numbers.

characteristics from the BRATS2015 dataset. However, it seems that our choice of the epoch number (i.e., 150) still works well. Certainly, 150 may not always work for any new dataset. In this case, we can again monitor the evolution of training error with respect to the number of epochs and choose the number for which training error becomes stable. In addition, as can be seen from Figure 4.7, within a broad range of epoch numbers, the test error does not change obviously. This indicates that the performance of our model may not be very sensitive to the choice of epoch number, and this also helps to relatively easily make a reasonable choice of this epoch number.

4.3.7 Generic Image Synthesis Results

Table 4.8: PSNR, NMSE, and SSIM on the generic image synthesis datasets (mean ± standard deviation). The paired t-test is conducted between dEa-GAN and Pix2pix to the significance level of 0.05. When the improvement of dEa-GAN is statistically significant, the result of Pix2pix will be underlined.

Methods	**facades**		
	PSNR	NMSE	SSIM
Pix2pix [30]	13.21±1.71	0.993±0.056	0.246±0.079
Proposed 2D dEa-GAN	**13.36±1.67**	**0.984±0.061**	**0.260±0.083**
Methods	**maps**		
	PSNR	NMSE	SSIM
Pix2pix [30]	15.06±2.08	0.878±0.065	0.203±0.083
Proposed 2D dEa-GAN	**15.60±2.09**	**0.870±0.068**	**0.237±0.083**
Methods	**cityscapes**		
	PSNR	NMSE	SSIM
Pix2pix [30]	15.98±2.41	0.851±0.097	0.421±0.083
Proposed 2D dEa-GAN	**16.46±2.55**	**0.844±0.100**	**0.435±0.086**

To evaluate the generality and the effectiveness of our edge-aware approach, we extend the dEa-GAN into its 2D variant, 2D dEa-GAN, and compare it with Pix2pix [30]. Following the literature, three generic image-to-image translation benchmark datasets are used. For the facades dataset [146], the label-to-photo translation is conducted with 400 training samples and 206 test samples. For the maps dataset, 1096 training images and 1098 test images that are scraped by [30] are used, and the map-to-aerial translation is processed. For the cityscapes dataset [147], photos are synthesized from the cityscapes labels with 2975 training images and 500 test images. All the image pre-processing step and experimental setting follow the work in [30].

In Table. 4.8, we report the quantitative comparisons of the synthesized images by Pix2pix [30] and our 2D dEa-GAN model on these three datasets. As consistently seen, our 2D dEa-GAN significantly outperforms Pix2pix [30] with higher PSNR and SSIM,

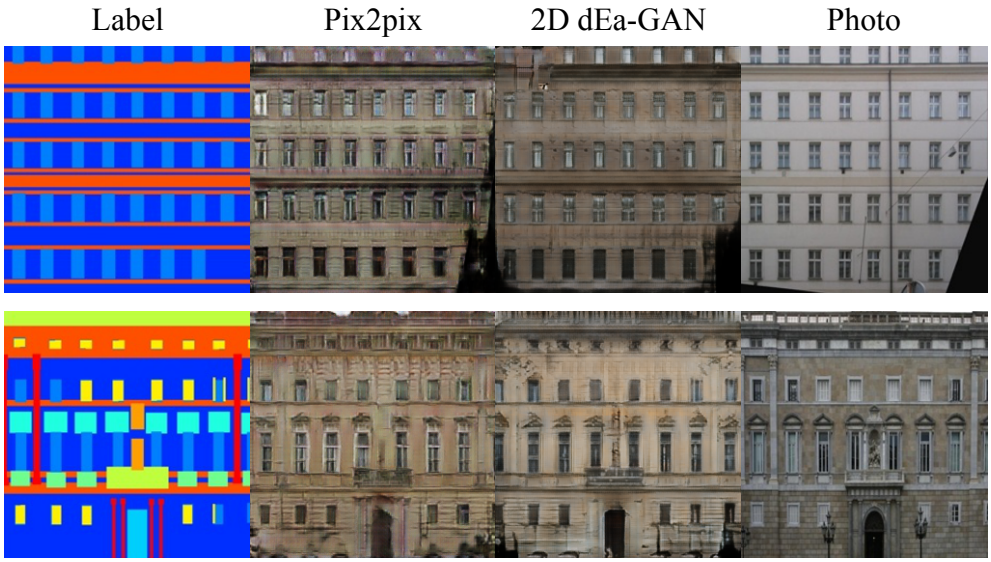


Figure 4.8: Comparison between Pix2pix and proposed 2D dEa-GAN on the facades dataset. The first example is in the top row, and the second example is presented in the bottom row.

Figure 4.9: Comparison between Pix2pix and proposed 2D dEa-GAN on the maps dataset. The first example is in the top row, and the second example is presented in the bottom row.

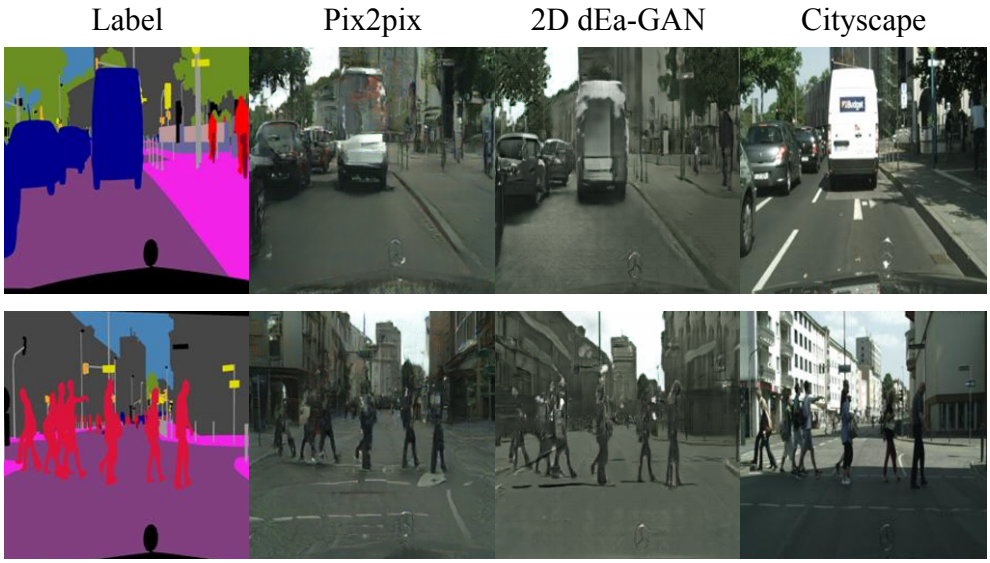


Figure 4.10: Comparison between Pix2pix and proposed 2D dEa-GAN on the cityscapes dataset. The first example is in the top row, and the second example is presented in the bottom row.

and lower NMSE. It validates that preserving edge information is essential for the different generic image synthesis tasks. For each dataset, visual comparisons of two examples are shown in Figures 4.8, 4.9, and 4.10, respectively. When looking into Figure 4.8, the surface of the outer wall and the contour of the door generated by 2D dEa-GAN is more intact than those estimated by Pix2pix. In Figure 4.9, our 2D dEa-GAN produces the clearer roofs of buildings and less blurred contours of rivers, compared with Pix2pix. Similarly, in Figure 4.10, cars and pedestrians are synthesized with the better appearance by the proposed 2D dEa-GAN. Therefore, both the visual and the quantitative results verify the generality and the capacity of our edge-aware GAN model in synthesizing generic images.

4.4 Discussion

This chapter aims to synthesize high-quality MR images by cGAN-based models. The proposed 3D-based Ea-GANs enforce the voxel-wise intensity similarity during training, and additionally integrate the edge maps as the image contextual information to improve synthesis performance. Two strategies have been proposed for this purpose. The first one is the proposed gEa-GAN model which extracts the Sobel edge maps of both synthesized and real target-modality images, and minimizes their distance during the training of generator. The second one is the proposed dEa-GAN. It further enforces this edge similarity via the adversarial learning between generator and discriminator. Our experimental results fully demonstrate the importance of perceiving the edge details during synthesis with the consistent improvements in terms of different evaluation measures, and across

all the datasets that have been tested. Moreover, by jointly acquiring the edge information via both of the generator and discriminator, the dEa-GAN is found to also outperform the proposed gEa-GAN that only incorporates edge details on the generator side. As for the computational time, in the gEa-GAN model, its number of trainable parameters is the same as that in the 3D cGAN model, so its training time remains almost unchanged. For the dEa-GAN model, using three input channels in the first layer of its discriminator only slightly increases the number of trainable parameters by 0.0065% (12,288 new parameters over totally 189,388,163 parameters). Therefore, using edge maps during the training incurs little computational burden.

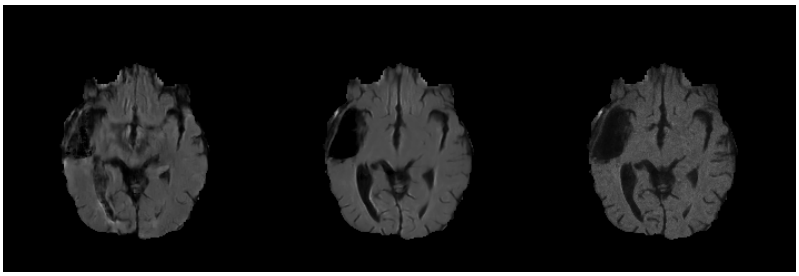
In addition, our Ea-GANs prove to also work well in lesion regions, which beats all the other methods in comparison according to three measurements on brain tumor synthesis. Last but not the least, edge-aware GANs well generalize to other generic image synthesis tasks, as shown on a variety of benchmark datasets about facades, maps, and cityscapes, demonstrating the power of our Ea-GAN model as a general image synthesizer.

More discussions about the model differences are presented in the following two sections.

4.4.1 Differences between the Existing cGANs and Ea-GANs

Most previous works for cGAN-based medical image synthesis [47, 55, 58, 60, 64, 66, 122, 125–129, 131–136] are 2D models that separately estimate each slice along the image's trans-axial direction. They ignore the 3D image contextual information and result in the discontinuous estimation. To overcome this limitation, the idea of using 3D conditional GANs has been exploited in some existing works [47, 129]. However, there are distinct differences between those methods and the proposed models. First, [129] only maintains the voxel-wise similarity during training. In contrast, our work innovatively explores the edge-aware idea in cGAN models to synthesize the higher-quality images. Second, whereas a gradient difference loss is used in [47], we apply a Sobel operator to extract the edge details. Applying the Sobel filters has more advantages than directly using the image gradients. By the averaging operation of Sobel operator, the filter is less sensitive to noise than directly using the image gradients. Also, the Sobel filter assigns higher weights to its nearer neighbors and lower weights to its farther neighbors, which cannot be achieved by directly utilising image gradients. The superiority of our gEa-GAN over the gradient cGAN justifies that the Sobel filters are more effective than the simple image gradients for MR image synthesis. More importantly, we innovatively integrate the edge information into adversarial learning (rather than simply putting it in the cost function of generator) to significantly improve the synthesis quality, which was not touched at all in [47].

4.4.2 Differences among the Compared Models



(a) Pix2pix (b) Multimodal (c) Replica

Figure 4.11: An example of the synthesized images by Pix2pix, Multimodal, and Replica

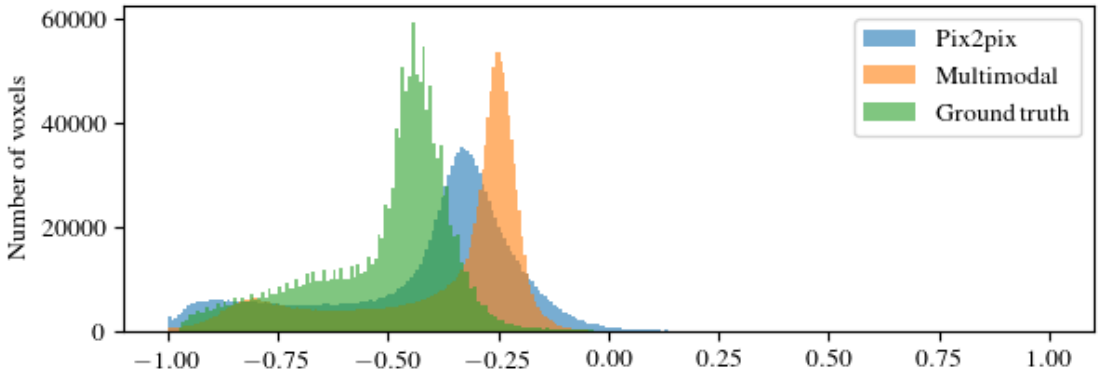


Figure 4.12: Histograms of the synthesized images by Pix2pix and Multimodal, and the ground truth

It is interesting to find that the coronal and sagittal views by Pix2pix show higher discontinuities than Replica and Multimodal in Figures 4.4, 4.5, and 4.6, although Pix2pix achieves better quantitative scores. The Replica is a 3D small-patch based method, so it produces the visually smooth results along all three directions. As shown in Figure 4.11, Pix2pix can synthesize sharper images than both Multimodal and Replica. Histograms of the two synthesized images by Pix2pix and Multimodal, as well as the real image are shown in Figure 4.12. We can see that, the intensity values of the image generated by Multimodal mainly vary over a small range, while those from Pix2pix spread over a larger range with larger distribution overlap to the ground truth. The larger variation of intensity values may be the reason of the relatively more salient discontinuity in the image produced by Pix2pix. However, it still produces the better quantitative results as evidenced by the more similar intensity distribution compared with the ground-truth.

4.5 Conclusion

In this chapter, we proposed two novel end-to-end learning 3D Ea-GANs models, i.e., gEa-GAN and dEa-GAN, to synthesize the target-modality MR images from the given paired source-modality images. Beyond the 3D cGAN model from Chapter 3 and other existing GAN synthesis models, the proposed Ea-GANs target to jointly preserve the voxel-wise intensity similarity and the edge similarity between the real and synthesized images to further preserve the crucial brain structure information in images. Our Ea-GANs, especially the more advanced dEa-GAN, achieve significantly better results than multiple state-of-the-art methods, i.e., the hand-crafted feature used model, the conventional CNNs, and the existing GANs, for MR image synthesis.

Chapter 5

Learning Sample-adaptive Local Sample Space Mappings for Cross-modality MR Image Synthesis

Most GANs attempt to train a unified model for all the input samples by learning a whole sample-space mapping from the source modality to the target modality. Whereas, compared with the complicated nature of voxel-wise prediction tasks, the available labeled samples of medical images are often scarce. Thus, it is challenging to optimize the unified model to fit all the samples via the global sample-space mapping. To mitigate this issue, this chapter develops a novel GANs based framework to learn both of the whole sample-space mapping and the local sample-space mapping by extracting the special characteristic of each individual sample for cross-modality lesion contained MR image synthesis tasks. To be specific, the developed framework decomposes the synthesis learning into a baseline path and an additional sample-adaptive path. The baseline path trains a common GAN model to conform to all the labeled samples as usual. The sample-adaptive path models each sample through its learnt relationship to its neighboring training samples and exploits their target-modality features as supplementary information during synthesis. Benefit from the cooperation between these two paths, the GAN model in the proposed synthesis framework is adaptive to the different input samples to increase the final synthesis results.

5.1 Motivation

The existing GANs models for cross-modality MR image synthesis have a main learning problem. They focus on learning a global mapping from the entire source-modality space to the whole target-modality space. Such a global sample space mapping can be highly complicated and hard to learn only from a small number of labeled medical samples.

This makes it difficult to achieve a global model that is optimal for every sample. Instead, we argue that the GAN models should be able to adapt to different samples to optimize the individual synthesis while at the same time conforming to the global sample space mapping mentioned above. Besides, in the existing GAN models, the target-modality is only used in the error term of an objective function for error evaluation. It is not actively sought to provide target-modality information to directly help the synthesis. For example, to synthesize FLAIR from T1, the target-modality FLAIR contains the valuable diffusion patterns around tumor regions, which cannot be clearly observed in the T1 [14]. Such target-modality information should be adequately utilized for synthesis.

To deal with these two issues, this chapter proposes sample-adaptive GAN models for cross-modality MR image synthesis. Essentially, a mapping of high-quality synthesis requires to extract the features that best represent the transformation from the source-modality to the target-modality samples. Therefore, our sample-adaptive GAN models consist of the learning of both common features and sample-specific features, and they are designed to cooperate with each other to boost the synthesis. Specifically, our sample-adaptive GANs comprise of two cooperative paths. The baseline path learns the common features for the global sample space mapping using a usual GANs model (backbone network), and the sample-adaptive path learns sample-specific characteristics by considering the local sample space details of each individual sample in its neighborhood. The two paths are jointly learned in an end-to-end manner. In this way, the proposed sample-adaptive GANs (SA-GANs) models can enforce the sample-specific learning on top of the common global sample space learning. Therefore, the unique characteristic of each input sample can be sample-adaptively learnt as the specific features and dynamically fit the synthesis task. Moreover, the training samples in the target modality are actively exploited to provide the auxiliary information to help the synthesis in both training and test stages. The effectiveness of our sample-adaptive models is validated on two lesion contained MR image datasets. The experimental results show the superior performance of our models over a set of state-of-the-art cross-modality image synthesis methods.

5.2 Proposed SA-GANs

5.2.1 Overview

The conventional deep learning based cross-modality image synthesis models learn a global sample space mapping and uniformly apply it to all samples in the whole space for prediction. However, due to the complexity of this global sample space mapping for synthesis and the scarcity of labeled training samples in most medical applications, it is very challenging to learn an optimal global model and extract powerful features representing the synthesis mapping. To deal with this issue, we propose novel sample-adaptive

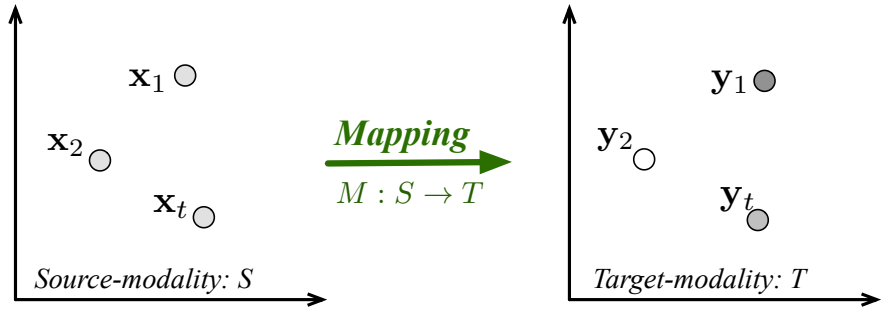
models, i.e., sample-adaptive GANs, to decompose the whole learning procedure into two cooperative paths, i.e., a baseline path and a sample-adaptive path. The baseline path learns the global sample space mapping and is trained by using all the training samples as usual, while the sample-adaptive path learns the sample-specific features via the unique characteristics of each individual sample. These two paths are jointly learned to provide complementary information to cater for the higher-quality synthesis in the proposed sample-adaptive GANs.

To best present the basic idea of the proposed SA-GANs, Figure 5.1 illustrates the comparison between the global and local sample space mappings. As given in Figure 5.1 (a), the global sample space mapping refers to a mapping M that maps the whole sample space in the source-modality to the whole sample space in the target-modality. This mapping is uniformly applied to every source-modality sample. Differently, a local sample space mapping, as presented in Figure 5.1 (b), maps a local sample region in the source-modality to a local sample region in the target-modality. Therefore, different samples could be mapped by a different mapping function between the two modalities. Specifically, the sample-adaptive path in our models will utilize K-nearest neighbors (KNNs) to build the local neighborhood around each sample and further use the generated weights of the neighbors to adjust the learned local sample space mapping. In this way, the sample-adaptive path can learn the specific local mappings for different input samples.

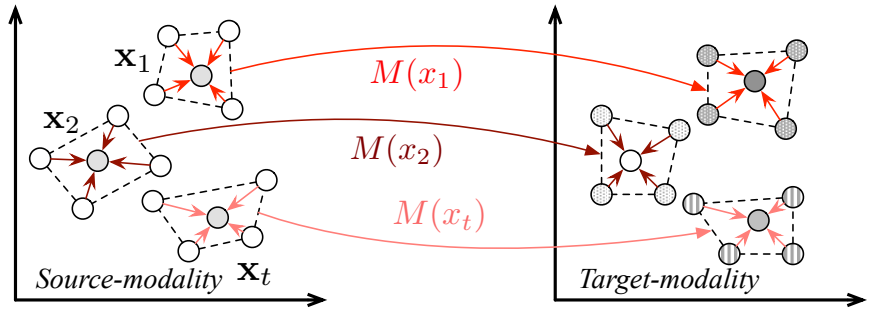
The detailed framework of the proposed sample-adaptive GANs is illustrated in Figure 5.2. As shown in the left part of Figure 5.2, the baseline path contains a GAN model, which extracts the common features by an embedding function $G^c(\cdot)$ with the parameter $\mathbf{W}^c$ (the superscript c means "common"). After that, it concatenates these common features with the sample-specific features learned by the sample-adaptive path for final prediction via another embedding function $G^g(\cdot)$ with the parameter $\mathbf{W}^g$ (the superscript g means final "generation"). In this common GAN model, both the trainable parameters, i.e., $\mathbf{W}^c$ and $\mathbf{W}^g$, are uniformly applied to all samples to achieve the global sample space mapping. In the right part of Figure 5.2, the sample-adaptive path learns the sample-specific features by an embedding function $F^s(\cdot)$ with the parameter $\mathbf{W}^s$ (the superscript s means "sample-specific"). Unlike the parameters in the baseline path, this parameter $\mathbf{W}^s$ is trained to capture the individual information from an input sample. The learned sample-specific features are then integrated into the baseline path as the auxiliary information to help the final synthesis. In the proposed sample-adaptive GANs, the baseline and the sample-adaptive paths are trained together in an end-to-end manner.

5.2.2 Baseline Path

In the baseline path, a usual 3D GAN model is used to learn the global sample space mapping from the source-modality to the target-modality. The generator in this GAN



(a) Global sample space mapping
(uniformly applied to the whole space)



(b) Local sample space mapping
(varied with different regions of the space)

Figure 5.1: Illustration of global sample space mapping and local sample space mapping.

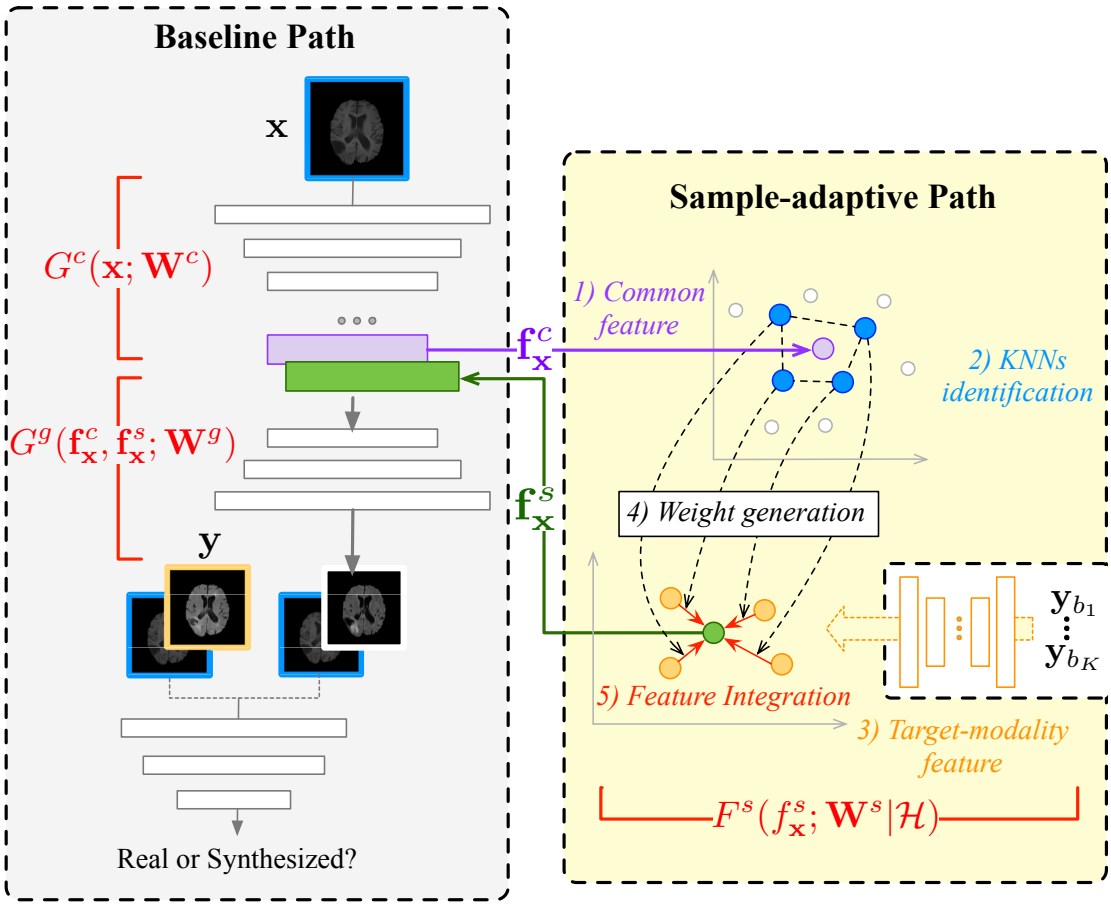


Figure 5.2: Framework of the proposed sample-adaptive GANs: a baseline path and a sample-adaptive path. The symbol $\mathscr{H}$ denotes the training set consisting of all source-modality samples $\mathbf{x}$ and their target-modality counterparts $\mathbf{y}$.

model, including a number of convolutional layers, aims to synthesize the realistic target-modality samples, while the discriminator D learns to differentiate the synthesized samples from the ground-truth. As mentioned in Section 5.2.1, the generator in this baseline path learns two embedding functions. Given an input sample $\mathbf{x}$, the embedding function $G^c(\cdot)$ covers from the input layer to the $(l-1)$-th layer in the generator to learn the common feature $\mathbf{f}_\mathbf{x}^c$. The embedding function $G^g(\cdot)$ covers from the l-th layer to the last output layer of the generator. It uses the concatenation of the common feature $\mathbf{f}_\mathbf{x}^c$ and the sample-specific feature $\mathbf{f}_\mathbf{x}^s$ (extracted from the sample-adaptive path to be introduced in Section 5.2.3) as input and outputs the final synthesis. In addition, as shown by the purple line in Figure 5.2, the common feature $\mathbf{f}_\mathbf{x}^c$ is also passed to the sample-adaptive path to guide the learning of the sample-specific feature $\mathbf{f}_\mathbf{x}^s$ for $\mathbf{x}$, which will be explained in detail in the following section. Consequently, the baseline and the sample-adaptive paths could be jointly trained through the above cooperation between them.

5.2.3 Sample-adaptive Path

Our sample-adaptive path is designed to estimate the unique feature of the corresponding target-modality sample based on an input source-modality sample. In other words, this feature should capture the individual characteristics of its counterpart in the target-modality space during the synthesis, and it is also called the target-modality feature in this chapter. Ideally, such a feature ought to be directly extracted from the real target-modality sample. However, this is infeasible because a real target-modality counterpart will not be available in the test stage. Therefore, in the sample-adaptive path, we propose to train a new network, which corresponds to the aforementioned embedding function $F^s(\cdot)$, to handle this issue. When given an input source-modality sample $\mathbf{x}$, this network learns to estimate the feature of its target-modality counterpart $\mathbf{y}$ by exploring its neighbors among the training samples. We assume that, given two samples $\mathbf{x}_i$ and $\mathbf{x}_j$, when they are close to each other (i.e., their features $f^c_{\mathbf{x}_i}$ and $f^c_{\mathbf{x}_j}$ are close), their target modality samples $\mathbf{y}_i$ and $\mathbf{y}_j$ shall also be close to each other. Mathematically, this assumes that the mapping from the source-modality space to the target-modality space is locally smooth. In this way, after identifying $\mathbf{x}$'s neighborhood through calculating the feature distance, the target-modality features of these neighboring training samples could be linearly combined to estimate $\mathbf{y}$'s feature. Here the target-modality features of these neighbors are extracted from their target-modality counterparts (which can be readily identified through the training set $\mathscr{H}$) by an auto-encoder trained externally using all the target-modality training samples, which will become clear shortly. To accommodate the different importance of each sample in this local sample space (neighborhood), a set of combination weights are predicted for these samples. Given the features of the sample $\mathbf{x}$ and its neighbors as the input, these weights are adaptively output by a weighting network whose parameters are jointly trained with the whole synthesis model. In this way, each sample has its unique sample-specific combination weights to depict the local space around it. Specifically, the input of this sample-adaptive path contains a given source-modality sample $\mathbf{x}$ and the entire training set including the paired source- and the target-modality training samples $\mathscr{H} = \{(\mathbf{x}_t, \mathbf{y}_t) \text{ for } t = 1, \ldots, T\}$. The symbol T denotes the entire number of training samples. To better present the learning procedure, the whole sample-adaptive path is decomposed into the following five steps in both training and test stages as shown in the right part of Figure 5.2.

1) Common Feature Extraction

As mentioned, the common feature $f^c_{\mathbf{x}}$ of $\mathbf{x}$ could be extracted by the first part of the generator, i.e., $G^c(\cdot)$. It is learned in the baseline path and applied to produce the features of all the source-modality training samples $\mathbf{x}_1, \ldots, \mathbf{x}_T$, which are denoted by $\mathbf{f}^c_{\mathbf{x}_1}, \ldots, \mathbf{f}^c_{\mathbf{x}_T}$.

2) K-nearest Neighbors Identification

The features extracted at Step 1) of common feature extraction are then used to identify the neighboring training samples for the input $\mathbf{x}$. After operating global average pooling on the feature $\mathbf{f}^c_{\mathbf{x}}$ of $\mathbf{x}$ and the feature of every training sample $\mathbf{f}^c_{\mathbf{x}_t}$ $(t = 1,\ldots,T, \text{and } \mathbf{x}_t \neq \mathbf{x})$, and then calculating their Euclidean distance, the KNNs of $\mathbf{x}$ are found and denoted by $\mathbf{x}_{b_1},\ldots,\mathbf{x}_{b_K}$. The features of these K neighbors are accordingly denoted by $\mathbf{f}^c_{\mathbf{x}_{b_1}},\ldots,\mathbf{f}^c_{\mathbf{x}_{b_K}}$.

3) Target-modality Feature Extraction

In the third step, a 3D CNN based auto-encoder model A, which is an external network apart from our sample-adaptive GANs, is used to extract the target-modality features for these K neighbors. This model is pre-trained by using all the target-modality training samples $\mathbf{y}_1,\ldots,\mathbf{y}_T$. Since the K neighbors of $\mathbf{x}$ are identified in Step 2), their target-modality counterparts can easily be known (i.e., from training pairs of $(\mathbf{x}_{b_k},\mathbf{y}_{b_k})$) and they are denoted by $\mathbf{y}_{b_1},\ldots,\mathbf{y}_{b_K}$. The pre-trained auto-encoder produces their target-modality features as $A(\mathbf{y}_{b_1}),\ldots,A(\mathbf{y}_{b_K})$. The detailed architecture of A will be introduced in Section 5.2.5.

4) Weight Generation

In this step, the sample-specific combination weights are generated in order to approximate the real target-modality feature of $\mathbf{x}$ from the target-modality features of $\mathbf{x}$'s K neighbors, which will be explained in Equation (5.3). The features $\mathbf{f}^c_{\mathbf{x}}$ and $\mathbf{f}^c_{\mathbf{x}_{b_1}},\ldots,\mathbf{f}^c_{\mathbf{x}_{b_K}}$ identified in Step 2) are input into a CNN based weighting network P to generate the combination weights through the end-to-end learning of our sample-adaptive GANs. These generated combination weights, which are denoted by $\omega_{b_1},\ldots,\omega_{b_K}$, reflect the importance of each neighbor in this combination. The architecture of P is presented in section 5.2.5.

5) Feature Integration

After the target-modality feature extraction at Step 3) and the weight generation at Step 4), we linearly combine the K target-modality features of neighbors with their corresponding combination weights. The sample-specific target-modality feature of the input sample $\mathbf{x}$ is estimated as $\mathbf{f}^s_{\mathbf{x}} = \sum_{k=1}^{K} \omega_{b_k} A(\mathbf{y}_{b_k})$. Then, this sample-specific feature $\mathbf{f}^s_{\mathbf{x}}$ is passed back to the baseline path to help the subsequent synthesis learning.

5.2.4 Objective Functions

The proposed sample-adaptive GANs have three sub-networks that need to be trained. The first and the second sub-networks are the discriminator D and the two parts of the generator, i.e., G^c and G^g, in the baseline path. The third sub-network is the weighting

network P in the sample-adaptive path and is trained to output the sample-specific feature $\mathbf{f}^s_{\mathbf{x}}$. When given a source-modality sample $\mathbf{x}$ and its real target-modality counterpart $\mathbf{y}$, i.e., the ground-truth, the two-player game between the generator and the discriminator in our sample-adaptive GANs follows the adversarial loss by training all the three sub-networks:

$$\begin{aligned}\mathscr{L}_{GAN}(G^c, G^g, D, P) = \mathbb{E}[\log D(\mathbf{x}, \mathbf{y})] + \\ \mathbb{E}[\log(1 - D(\mathbf{x}, G^g(G^c(\mathbf{x}), \mathbf{f}^s_{\mathbf{x}})))],\end{aligned} \tag{5.1}$$

where, as mentioned before, $G^c(\cdot)$, $G^g(\cdot)$, and $D(\cdot)$ denote the outputs of the two parts in the generator and the discriminator, respectively, and $\mathbb{E}$ indicates mathematical expectation.

Additionally, an L1-norm penalty on the voxel-wise intensity difference between the synthesized target-modality sample $G^g(G^c(\mathbf{x}), \mathbf{f}^s_{\mathbf{x}})$ and the ground-truth $\mathbf{y}$ is applied to enforce their similarity via the following objective function:

$$\mathscr{L}_{sample}(G^c, G^g, P) = \mathbb{E}[\|\mathbf{y} - G^g(G^c(\mathbf{x}), \mathbf{f}^s_{\mathbf{x}}))\|_1]. \tag{5.2}$$

Moreover, in the sample-adaptive path, in order to make the learned feature $\mathbf{f}^s_{\mathbf{x}}$ to approximate the ground-truth target-modality feature $A(\mathbf{y})$, the weighting network P is trained (via the parameter $\mathbf{W}^s$) to best generate the combination weights $\omega_{b_1}, \ldots, \omega_{b_K}$ by minimizing the following objective function:

$$\mathscr{L}_{feature}(P, G^c) = \mathbb{E}[\|A(\mathbf{y}) - \mathbf{f}^s_{\mathbf{x}}\|_2^2]. \tag{5.3}$$

Note that, since the neighbors are identified by using $\mathbf{f}^c_{\mathbf{x}}$, and our GANs models are trained end-to-end, optimizing Equation (5.3) will also contribute to the learning of $G^c(\cdot)$.

To synthesize realistic target-modality samples, the three sub-networks of our sample-adaptive GANs are jointly trained by integrating the above three objective functions into a final one as follows:

$$\begin{aligned}\mathscr{L}_{final} = \arg\min_{G^c}\min_{G^g}\max_{D}\min_{P}\mathscr{L}_{GAN}(G^c, G^g, D, P) + \\ \lambda\mathscr{L}_{sample}(G^c, G^g, P) + \zeta\mathscr{L}_{feature}(P, G^c),\end{aligned} \tag{5.4}$$

where both λ and ζ are the hyper-parameters to balance these terms. In Algorithm 1, the training details of updating the proposed SA-GANs are presented.

Finally, to achieve better synthesis performance, a more delicate sample-specific feature learning is practically implemented. Specifically, each local spatial part of an input image sample has its unique linear combination. When the feature size of the entire sample is $C \times N \times N \times N$ (C is the number of feature channels), the target-modality feature of

Algorithm 1 Training SA-GANs

input: The entire training set $\mathscr{H} = \{(\mathbf{x}_t, \mathbf{y}_t) \text{ for } t = 1, \ldots, T\}$, the number of training epochs N_{epoch}, and two hyper-parameters λ and ζ.

while $n_e < N_{epoch}$ **do**
  for $(\mathbf{x}, \mathbf{y})$ in $\mathscr{H}$ **do**
    Get $G^c(\mathbf{x})$ by G^c from $\mathbf{x}$.
    Get $\mathbf{f}^s_{\mathbf{x}}$ by P from $G^c(\mathbf{x})$ and $\{(\mathbf{x}_t, \mathbf{y}_t)$ for $t = 1, \ldots, T$, and $\mathbf{x}_t \neq \mathbf{x}\}$.
    Get $G^g(G^c(\mathbf{x}), \mathbf{f}^s_{\mathbf{x}})$ by G^g.
    Get $D(\mathbf{x}, \mathbf{y})$ and $D(\mathbf{x}, G^g(G^c(\mathbf{x}), \mathbf{f}^s_{\mathbf{x}})))$ by D.
    $\mathscr{L}_{GAN} \leftarrow \log D(\mathbf{x}, \mathbf{y}) + \log(1 - D(\mathbf{x}, G^g(G^c(\mathbf{x}), \mathbf{f}^s_{\mathbf{x}})))$.
    Maximize $\mathscr{L}_{GAN}$ to update D.
    $\mathscr{L}_{sample} \leftarrow \|\mathbf{y} - G^g(G^c(\mathbf{x}), \mathbf{f}^s_{\mathbf{x}}))\|_1$.
    $\mathscr{L}_{feature} \leftarrow \|A(\mathbf{y}) - \mathbf{f}^s_{\mathbf{x}}\|_2^2$.
    $\mathscr{L}_{final} \leftarrow \mathscr{L}_{GAN} + \lambda \mathscr{L}_{sample} + \zeta \mathscr{L}_{feature}$.
    Minimize $\mathscr{L}_{final}$ to update G^c, G^g, and P together.
  end for
  $n_e = n_e + 1$
end while

output: The trained two parts of generator G^c and G^g, weighting network P and the discriminator D.

this local part (size: $C \times M \times M \times M$) is obtained by integrating the target-modality features of its own K nearest small spatial parts with their corresponding local combination weights. Therefore, each input sample $\mathbf{x}$ has $(N/M)^3$ locally sample-specific combination ways to depict the spatial characteristics in our end-to-end trained sample-adaptive path.

5.2.5 Network Architectures

Weighting Network

As mentioned, the sample-specific feature learning is spatially applied to the features of each small part. In this work, the size (M) of these features is set as four to achieve a more delicate feature learning. Every local part of K neighbors has its unique corresponding weight. Thus, a three-convolutional-layer weighting network P, as illustrated in Figure 5.3, is designed to generate K combination weights for the K neighbors at the same time. The detailed architecture of the weighting network is as follows: (1) a conv layer with $2 \times 2 \times 2$ kernel size, 2 stride, and 256 output channels (the number of input channels: $C * (K + 1)$), a BN layer, and a ReLU layer, (2) a conv layer with $2 \times 2 \times 2$ kernel size, 2 stride, and 64 output channels, a BN layer, and a ReLU layer, and (3) a conv layer with $1 \times 1 \times 1$ kernel size, 1 stride, and K output channels, and a Sigmoid layer. With the Sigmoid layer, the values of all these K generated combination weights are in $(0, 1)$.

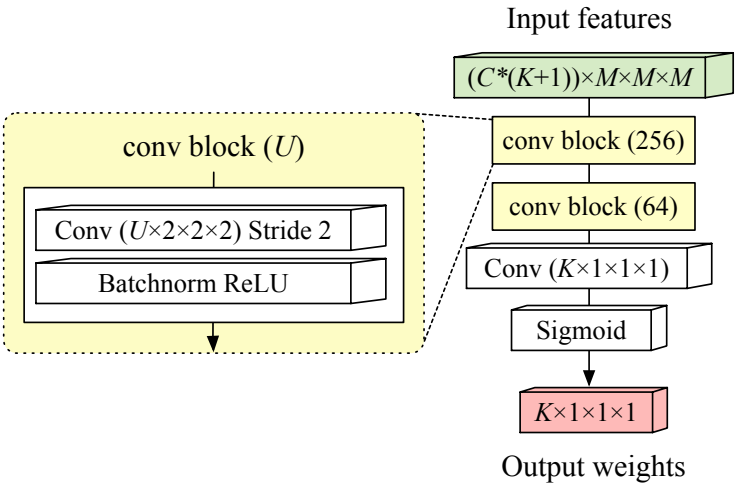


Figure 5.3: Architecture of weighting network. The weighting network contains three convolutional layers to generate K combination weights with their values in $(0, 1)$.

3D GAN

In this chapter, we integrate the proposed sample adaptive strategy with two backbone GANs models in order to demonstrate the flexibility of our strategy to different GANs. These two models are the 3D cGAN model and the dEa-GAN model that are presented in Chapter 3. Using these two models to construct the common path respectively, we call our resulting models SA-GAN (with respect to 3D cGAN) and dEA-SA-GAN (with respect to dEa-GAN) accordingly. The generators of both models take the U-net-like architecture including seven convolutional layers and seven up-convolutional layers with the skip-connections between each pair of them, and their discriminators are CNN-based networks with six convolutional layers. We follow their original architectures except that in the l-th layer of their generators, the number of input channels is increased due to the additional C channels of the sample-specific feature learned from the sample-adaptive path. When the size of input samples is $128 \times 128 \times 128$, the number of l is 11. When the input size is $64 \times 64 \times 64$, we remove the second to last convolutional layer and the second up-convolutional layer from the original generators of both SA-GAN and dEa-SA-GAN models to fit the input size and therefore set the number of l as nine.

Auto-encoder

The auto-encoder A is an external pre-trained model used to extract the target-modality features for the sample-adaptive path. It contains an encoder and a decoder, as shown in Figure 5.4. The encoder consists of four convolutional (conv) blocks, and the decoder includes four up-convolutional (up-conv) blocks. Each of these eight conv and up-conv blocks has three layers, orderly corresponding to a conv (or up-conv) layer with $4 \times 4 \times 4$ kernel size and 2 stride (or 1/2 stride for up-conv layer), a batch normalization (BN)

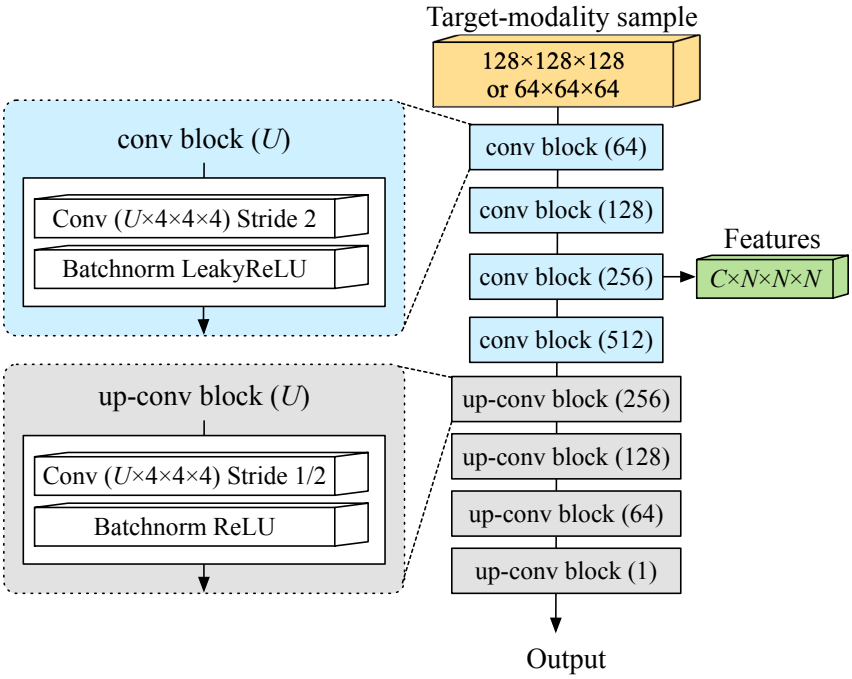


Figure 5.4: Architecture of auto-encoder. The auto-encoder includes a four-convolutional-layer encoder and a four-up-convolution-layer decoder. When it is applied to extract target-modality features, these features are produced from the third convolutional block in its encoder.

layer, and a ReLU layer. The ReLU layers in the encoder are LeakyReLU with the slope of 0.2. From the shallower to the deeper layers of the auto-encoder, the numbers of output channels in these eight blocks are 64, 128, 256, 512, 256, 128, 64, and 1, respectively. When the size of input samples is $128 \times 128 \times 128$ or $64 \times 64 \times 64$, the target-modality features $A(\cdot)$ extracted from the third conv block in the encoder will have the size of $256 \times 16 \times 16 \times 16$ ($C = 256$ and $N = 16$) or $256 \times 8 \times 8 \times 8$ ($C = 256$ and $N = 8$).

5.2.6 Implementation

As discussed in [141], training GAN models experience a common issue that they may become unstable. This is due to that the discriminator more easily has the powerful ability than the generator, which leads to the unbalanced competition between these two agents during training. At the initial stage of training the proposed sample-adaptive GANs, the extracted common features from the first part of their generators, i.e., G^c, tend to be of poor quality. This may degrade the accuracy of KNNs identification which is an important step in the sample-adaptive path. As a result, it will adversely affect the performance of the subsequent feature integration and final synthesis, and even negatively impact the balance between the generator and the discriminator. To deal with this issue, a training strategy is utilized to better improve the synthesis performance in the initial stage by gradually increasing the difficulty of feature learning in the sample-adaptive path. Specifically,

in the first 20 training epochs, the real target-modality feature $A(\mathbf{y})$ rather than the learned source-modality feature is applied to directly identify its neighbors. By this means, the parameters in the entire proposed models including those in G^c could be well trained at the beginning. After these 20 epochs, G^c could estimate better source-modality features which will be used in the KNNs identification as presented in Section 5.2.3. Therefore, this strategy could raise the robustness of the proposed GAN models during training.

Deep learning models, especially the 3D CNN based ones, contain numerous trainable parameters to capture the comprehensive and the hidden information from the input. As a result, compared with the conventional models which only use hand-crafted features, training a deep learning model requires a much longer period of time. For the proposed SA-GAN and dEa-SA-GAN models, all the training samples should be passed through G^c once to extract updated features for the input samples in every training batch. This will further increase the training time. In this case, it is a trade-off between training efficiency and training performance. Thus, we extract the features $\mathbf{f}^c_{\mathbf{x}_t}$ $(t = 1, \ldots, T, \text{and } \mathbf{x} \neq \mathbf{x}_t)$ of all the training samples except $\mathbf{x}$ once for every 30 training batches during the sample-specific feature learning. This strategy could significantly improve training efficiency but have little negative impact on the training quality.

5.3 Experimental Results

5.3.1 Datasets

The proposed SA-GAN and dEa-SA-GAN models are evaluated on two MRI datasets, i.e., brain tumor contained BRATS2015 [14] and stroke lesion contained SISS2015 [16]. The details about BRATS2015 dataset has been presented in Chapter 4. The SISS2015 dataset has 28 subjects with the MR images (size: $230 \times 230 \times 153$ pixels or $230 \times 230 \times 154$ pixels) from T1w, T2w, FLAIR, and diffusion-weighted imaging (DWI) modalities. For this dataset, we conduct a synthesis task from T1 to FLAIR. Similar to Chapter 4, the proposed models are trained on the BRATS2015 and SSIS2015 datasets using 5-fold cross-validation (except from the experiments in ablation study and segmentation evaluation), and the original intensity values are re-scaled to [-1,1]. Since the SISS2015 dataset consists of fewer subjects than BRATS2015, we extract 48 large patches with size $64 \times 64 \times 64$ from each MR image, and also average the overlapped regions in the final estimation. All the MR images of the same subject in these two datasets have been co-registered.

5.3.2 Experimental Settings

Both SA-GAN and dEa-SA-GAN models are trained with 150 epochs for all three synthesis tasks. The learning rate of the three sub-networks in the proposed models is set as 0.0002 in the first 100 epochs, and then it linearly decays to zero in the last 50 epochs. Adam solver with the mini-batch size of six is applied to optimize the proposed models. The hyper-parameters λ and ζ in Equation. 5.4 are set as 300 and 150 to balance the terms in the training objectives, respectively. The number of searched nearest neighbors, i.e., K, in the sample-adaptive path is set as seven for the two tasks on the BRATS2015 dataset and five on the SISS2015 dataset considering training efficiency and computational cost.

5.3.3 Methods in Comparison

The proposed SA-GAN and dEa-SA-GAN models are compared with their corresponding backbone models, i.e., the original 3D cGAN and dEa-GAN from Chapter 4, and other six state-of-the-art cross-modality MR image synthesis models in the recent literature: Replica [41], Multimodal [63], pGAN [64], perceptual GAN [32], WGAN-GP [28], and gradient cGAN [148].

1. 3D cGAN from Chapter 3 is the backbone of the baseline path in SA-GAN. It learns a global space mapping for synthesis by a 3D GAN model with the objective to ensure the voxel-wise intensity similarity.

2. dEa-GAN from Chapter 4 is the backbone of the baseline path in dEa-SA-GAN. It preserves the edge information of 3D MR samples during learning a global space mapping for synthesis through adversarial learning.

3. Replica [41] is an approach that trains random forests with the hand-crafted features from multi-resolution 3D patches for synthesis.

4. Multimodal [63] learns a global space mapping to synthesize the axial slices of MR images by a conventional 2D CNN-based model.

5. pGAN [64] is a 2D GAN model, learning a global space mapping to synthesize the axial slices of MR images with the objective to ensure the pixel-wise intensity similarity.

6. Perceptual GAN [32] is based on the same 2D GAN model as the pGAN [64], but it additionally uses a pre-trained VGG encoder to extract the features of the real and the synthesized sample pairs and enforces their features to be similar through adversarial training. Perceptual GAN also learns a global sample space mapping like pGAN [64].

7. WGAN-GP [28] uses the same 2D GAN model as the pGAN [64], but it employs Wasserstein distance for adversarial learning to improve the stability of training procedure. It still learns a global sample space mapping.

8. gradient cGAN from Chapter 4 is a usual 3D GAN model and trained with an additional gradient similarity loss from [47].

The experiments on these seven compared models are conducted by using the codes obtained from their authors and following their original papers for both image pre-processing steps and model settings. Their codes are re-run on the two datasets employed in this work to ensure fair comparisons.

5.3.4 Ablation Study

Table 5.1: Ablation study on K by the proposed SA-GAN for three synthesis tasks according to PSNR values.

K	FLAIR (BRATS2015)	T2 (BRATS2015)	FLAIR (SISS2015)
$K = 3$	30.14	29.68	30.39
$K = 5$	30.19	29.56	30.47
$K = 7$	30.37	29.83	30.32
3D cGAN	29.21	28.98	29.45

In the proposed SA-GANs, the key hyper-parameter is the number of neighbors, i.e., K, used in the sample-adaptive path. To study its effect on the final synthesis performance, we conduct the sensitivity experiment by setting $K = 3, 5,$ and 7 on the three synthesis tasks, respectively. We randomly partition each dataset into 80% for training and 20% for test in this experiment. The PSNR values of their synthesis results are presented in Table 5.1. As shown, the performance of the proposed SA-GAN models with different Ks is not very sensitive to the change of the K value. Moreover, its performance is consistently higher than that of the baseline model, i.e., 3D cGAN. In this work, we set Ks as seven for all the experiments on BRATS2015 dataset (its size is relatively large, with 274 subjects) and five on SSIS2015 dataset (its size is relatively small, with 28 subjects).

5.3.5 Results on BRATS2015

In this section, we present the experimental results on the BRATS2015 dataset. The quantitative results of the two synthesis tasks on whole images including the brain region and the background are reported in the upper parts of Tables 5.2, 5.3 and 5.4. Moreover, the synthesis results evaluated on tumor regions are provided in the lower parts

of Tables 5.2, 5.3 and 5.4. In order to test the significance of the improvements from baselines, a paired t-test is applied. If the improvements of the proposed SA-GAN and dEa-SA-GAN models over their corresponding baselines, i.e., 3D cGAN and dEa-GAN, are statistically significant, the results from the proposed methods will be underlined in Tables 5.2 and 5.3. In this work, we use the significance level of 0.05.

Comparison with Baselines

To study the effectiveness of the proposed sample-adaptive learning strategy, we compare the SA-GAN and dEa-SA-GAN with their corresponding baselines, i.e., the original 3D cGAN and dEa-GAN. The results on whole images in two synthesis tasks are reported in the upper parts of Tables 5.2 and 5.3, respectively. As shown, both proposed methods SA-GAN and dEa-SA-GAN outperform their corresponding baselines by achieving higher PSNR and SSIM and lower NMSE consistently over the two synthesis tasks of T1-to-FLAIR and T1-to-T2.

The advantages of the proposed methods become more salient on the tumor regions, as reported in the lower parts of Tables 5.2 and 5.3. Specifically, SA-GAN improves 3D cGAN on tumor regions by about 1.12dB PSNR, 0.009 NMSE, and 0.028 SSIM in the T1-to-FLAIR task, and 1.11dB PSNR, 0.007 NMSE, and 0.040 SSIM in the T1-to-T2 task, respectively. Also, dEa-SA-GAN raises the performance of dEa-GAN by 1.44dB PSNR, 0.003 NMSE, and 0.040 SSIM in the T1-to-FLAIR task, and 1.18dB PSNR, 0.005 NMSE, and 0.029 SSIM in the T1-to-T2 task, respectively.

Please note that the relatively large standard deviation (STD) values are mainly due to the variation of images rather than purely from methods in comparison. Therefore, the paired t-test is employed to purely compare two methods, which calculates the STD of the *difference* of the compared results rather than the STD of the results directly. For example, in terms of PSNR on whole images, the STD of the *difference* between SA-GAN and 3D cGAN is only 0.79 and that between dEa-SA-GAN and dEa-GAN is 0.63, which are in contrast to the sample-based STD values that are around 3. According to the paired t-test, our improvements are statistically significant (at the level of 0.05) on both tumor regions and whole images.

To further explore the advantages of our proposed methods, the numbers of subjects that the proposed models win or lose to their corresponding baselines (in terms of PSNR) are also given in Tables 5.2 and 5.3. As shown, for more than 80% of all subjects in BRATS2015 dataset, our proposed models outperform their corresponding baselines for both whole image synthesis (except dEa-SA-GAN vs dEA-GAN on T1-to-FLAIR task. In that case, ours outperforms on 72% of all subjects.) and tumor region synthesis. These results are consistent with the results of the paired t-test, which reinforces the statistical significance of our improvements.

Table 5.2: T1 to FLAIR on BRATS2015 dataset: comparison with **baselines**. The evaluations (mean±STD) are provided for both the whole images and the tumor regions. The win/lose indicates the number of subjects that the proposed models win / lose to their corresponding baselines according to PSNR values.

	Methods	PSNR	NMSE	SSIM	win/lose
Whole image	3D cGAN from Chapter 3 (baseline 1)	29.26±3.21	0.119±0.205	0.958±0.016	236/38
	SA-GAN (proposed)	**30.21±3.10**	**0.104±0.181**	**0.962±0.015**	
	dEa-GAN from Chapter 4 (baseline 2)	30.11±3.22	0.105±0.174	0.963±0.016	197/77
	dEa-SA-GAN (proposed)	**30.68±3.14**	**0.101±0.178**	**0.965±0.016**	
Tumor region	3D cGAN from Chapter 3 (baseline 1)	15.95±3.52	0.098±0.094	0.681±0.090	229/45
	SA-GAN (proposed)	**17.07±3.38**	**0.089±0.097**	**0.709±0.089**	
	dEa-GAN from Chapter 4 (baseline 2)	16.90±3.59	0.084±0.099	0.705±0.093	242/32
	dEa-SA-GAN (proposed)	**18.34±3.47**	**0.081±0.099**	**0.745±0.100**	

Table 5.3: T1 to T2 on BRATS2015 dataset: comparison with baselines. The evaluations (mean±STD) are provided for both the whole images and the tumor regions. The win/lose indicates the number of subjects that the proposed models win / lose to their corresponding baselines according to PSNR values.

	Methods	PSNR	NMSE	SSIM	win/lose
Whole image	3D cGAN from Chapter 3 (baseline 1)	29.34±3.23	0.095±0.199	0.964±0.017	215/59
	SA-GAN (proposed)	**30.01±3.11**	**0.087±0.203**	**0.967±0.016**	
	dEa-GAN from Chapter 4 (baseline 2)	29.98±3.37	0.088±0.223	0.967±0.016	211/63
	dEa-SA-GAN (proposed)	**30.62±3.10**	**0.082±0.192**	**0.970±0.017**	
Tumor region	3D cGAN from Chapter 3 (baseline 1)	16.79±3.56	0.089±0.093	0.725±0.099	230/44
	SA-GAN (proposed)	**17.90±3.42**	**0.082±0.085**	**0.765±0.098**	
	dEa-GAN from Chapter 4 (baseline 2)	18.02±3.55	0.079±0.098	0.766±0.098	239/35
	dEa-SA-GAN (proposed)	**19.20±3.27**	**0.074±0.082**	**0.795±0.101**	

Table 5.4: Comparison with the state-of-the-art methods on the BRATS2015 dataset. Evaluations (mean±STD) are provided for both the whole images and the tumor regions.

	Methods	T1 to FLAIR			T1 to T2		
		PSNR	NMSE	SSIM	PSNR	NMSE	SSIM
Whole image	Replica [41]	27.17±2.60	0.171±0.267	0.939±0.013	26.92±2.36	0.158±0.324	0.946±0.015
	Multimodal [63]	27.26±2.82	0.184±0.284	0.950±0.014	27.31±2.39	0.140±0.229	0.951±0.016
	pGAN [64]	27.46±2.55	0.144±0.189	0.940±0.015	28.12±2.45	0.110±0.220	0.953±0.014
	perceptual GAN [32]	25.88±3.24	0.290±0.211	0.898±0.019	24.99±3.25	0.307±0.225	0.873±0.022
	WGAN-GP [28]	27.37±2.62	0.146±0.201	0.943±0.015	28.13±2.85	0.117±0.287	0.955±0.015
	gradient cGAN from Chapter 4	29.38±3.25	0.116±0.204	0.960±0.017	29.43±3.28	0.097±0.210	0.966±0.017
	SA-GAN (proposed)	**30.21±3.10**	**0.104±0.181**	**0.962±0.015**	**30.01±3.11**	**0.087±0.203**	**0.967±0.016**
	dEa-SA-GAN (proposed)	**30.68±3.14**	**0.101±0.178**	**0.965±0.016**	**30.62±3.10**	**0.082±0.192**	**0.970±0.017**
Tumor region	Replica [41]	13.34±3.41	0.137±0.068	0.601±0.083	14.93±3.17	0.123±0.084	0.650±0.139
	Multimodal [63]	13.82±3.66	0.131±0.076	0.638±0.096	15.50±3.75	0.109±0.117	0.689±0.138
	pGAN [64]	14.48±3.12	0.127±0.093	0.618±0.084	16.03±3.10	0.099±0.084	0.703±0.095
	perceptual GAN [32]	13.18±3.54	0.189±0.104	0.548±0.097	13.33±3.54	0.212±0.113	0.488±0.104
	WGAN-GP [28]	14.51±3.29	0.125±0.088	0.630±0.086	15.90±3.44	0.102±0.097	0.712±0.091
	gradient cGAN from Chapter 4	15.67±3.63	0.104±0.123	0.682±0.090	16.87±3.40	0.085±0.089	0.752±0.098
	SA-GAN (proposed)	**17.07±3.38**	**0.089±0.097**	**0.709±0.089**	**17.90±3.42**	**0.082±0.085**	**0.765±0.098**
	dEa-SA-GAN (proposed)	**18.34±3.47**	**0.081±0.099**	**0.745±0.100**	**19.20±3.27**	**0.074±0.082**	**0.795±0.101**

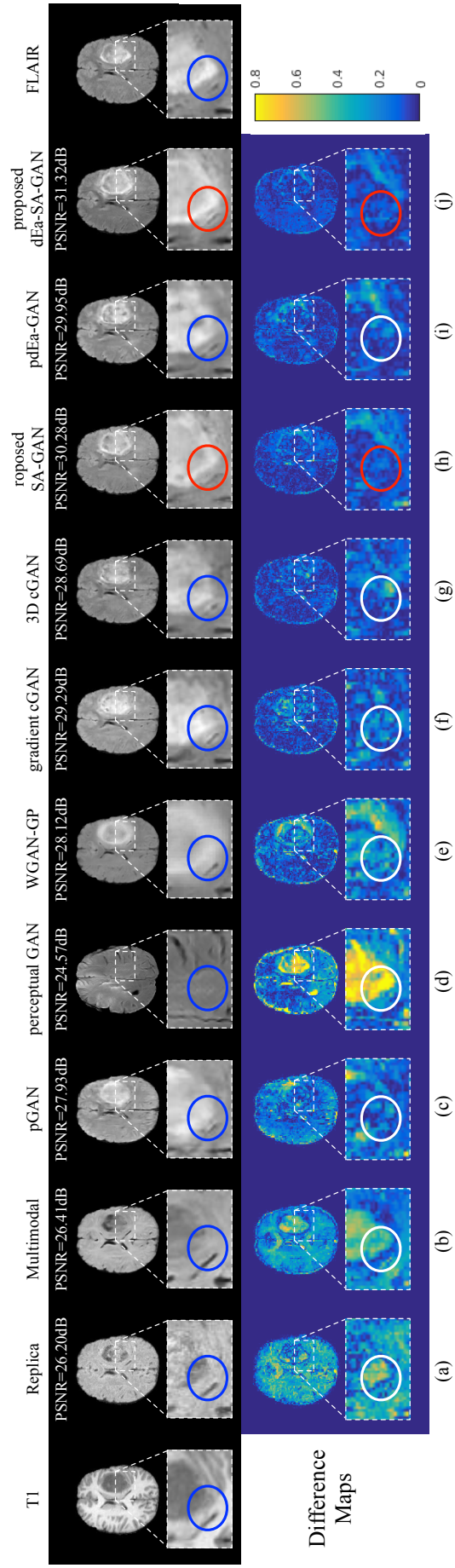


Figure 5.5: Comparison between the two proposed models and other state-of-the-art methods (T1 to FLAIR on BRATS2015 dataset). The (a)-(j) regions in circles contain the boundaries between tumor and healthy tissues. As shown by the small values in the difference maps of (h) and (j), they are better synthesized by SA-GAN and dEa-SA-GAN methods.

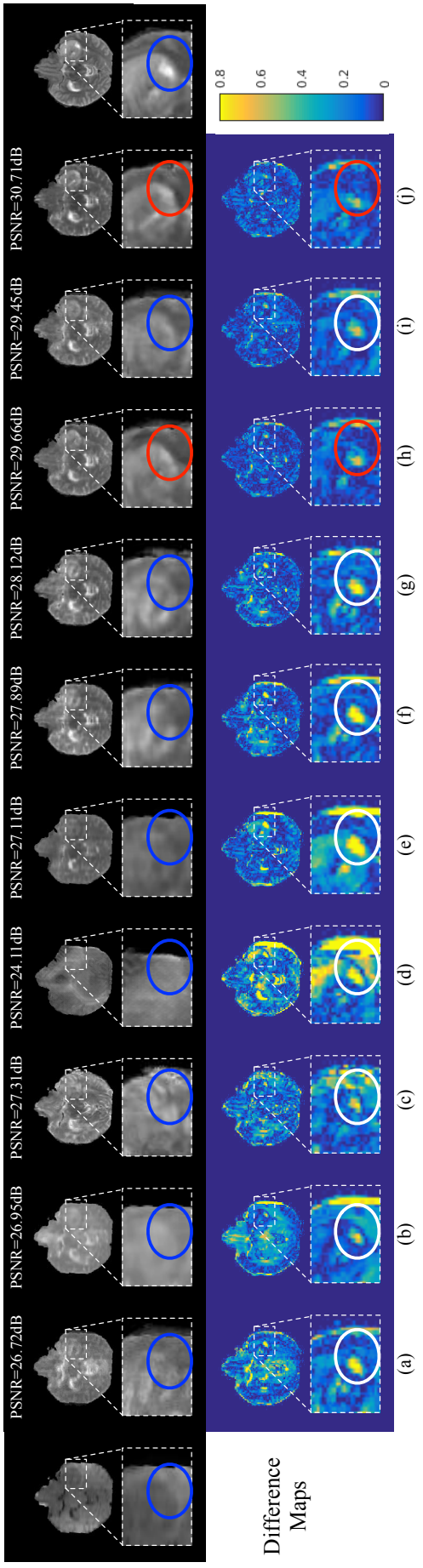


Figure 5.6: Comparison between the two proposed models and other state-of-the-art methods (T1 to T2 on BRATS2015 dataset). The (a)-(j) regions in circles contain the boundaries between tumor and healthy tissues. As shown by the small values in the difference maps of (h) and (j), they are more clearly synthesized by SA-GAN and dEa-SA-GAN methods.

Comparisons with State-of-the-art Methods

We further compare the proposed SA-GAN and dEa-SA-GAN with another six state-of-the-art models, i.e., Replica [41], Multimodal [63], pGAN [64], perceptual GAN [32], WGAN-GP [28], and gradient cGAN from Chapter 4 for cross-modality MR image synthesis. As shown in the upper part of Table 5.4, both SA-GAN and dEa-SA-GAN achieve better results than the other four compared methods in terms of all three evaluation measurements on both synthesis tasks. Among the compared methods, the perceptual GAN [32] seems to have inferior performance. This may be due to the fact that the VGG model pre-trained on generic image dataset (e.g., ImageNet) cannot be directly applied to medical images since they have a significant different nature from the generic images. The second worst results are obtained by Replica [41], indicating the importance of learning deep features and using whole-image/large-patches for synthesis. Also, except perceptual GAN [32], the other five GAN-based models perform better than Multimodal [63] using CNN, showing the advantages of adversarial learning. Further comparing the results from our proposed methods with those from pGAN [64] and gradient cGAN from Chapter 4, we could observe that SA-GAN and dEa-SA-GAN generate better synthesis with considerable improvements than those two GAN models using different training objectives and learning the global sample space mapping for synthesis. Since gradient cGAN from Chapter 4 has better performance in the four methods in comparison, paired t-tests are conducted between gradient cGAN and SA-GAN, and between gradient cGAN and dEa-SA-GAN. The results of the tests indicate that almost all the improvements of the proposed models from gradient cGAN are statistically significant (at the level of 0.05). Again, from the lower part of Table 5.4, we can see that the superiority of the proposed models over other compared methods become more salient on the synthesized tumor regions. This is similar to our observations in comparison with the baselines.

Two visual examples of the results are provided in Figure 5.5 and Figure 5.6. It can be seen that the proposed SA-GAN and dEa-SA-GAN produce visually clearer local details when the synthesized images are zoomed in the rectangles, and their generated images seem to have higher fidelity with the ground-truth.

Segmentation Evaluation on the Synthesized Images

In order to further study the quality of synthesized tumor parts, the popular brain tumor segmentation model Deepmedic [107] is applied to examine the segmentation performance obtained using the synthesized FLAIR and T2 images. We use two schemes to evaluate the dice scores of whole tumors segmented from the synthesized images produced by SA-GAN and dEa-SA-GAN, respectively, and compare them with the results from the two baselines, i.e., 3D cGAN and dEa-GAN. In the first scheme (denoted as Real&Syn), we trained the segmentation model using the real FLAIR or T2 images (i.e.,

Table 5.5: Dice scores (mean±STD) of whole tumor segmentation on the synthesized images. Paired t-tests are conducted between the proposed methods, i.e., SA-GAN and dEa-SA-GAN, and their baselines at the significance level of 0.05, respectively. When the segmentation improvement is statistically significant, the result of the proposed method will be underlined.

Methods	**Real&Syn**		**Syn&Syn**	
	FLAIR	T2	FLAIR	T2
3D cGAN	0.667±0.194	0.658±0.189	0.671±0.179	0.665±0.203
SA-GAN	**0.707±0.201**	**0.702±0.200**	**0.703±0.190**	**0.710±0.195**
dEa-GAN	0.689±0.174	0.698±0.180	0.694±0.182	0.688±0.180
dEa-SA-GAN	**0.721±0.173**	**0.711±0.186**	**0.715±0.179**	**0.718±0.183**

target modality) from training set (randomly selected 80% data). This model was then applied to segment the tumors in the synthesized FLAIR or T2 images from test set (the rest 20% data). The results are shown in the left part of Table 5.5. As can be seen, the segmentation model achieved significantly better results on the images synthesized by our sample-adaptive GANs (SA-GAN and dEa-SA-GAN) than those synthesized by the corresponding baselines (3D cGAN and dEa-GAN). This suggests that our synthesized FLAIR or T2 images more resemble the real target modality images (from the perspective of segmentation), so that the segmentation model trained using the real target modality images could be better generalized to our synthesized images. In the second scheme (denoted as Syn&Syn), we trained the segmentation models using the training set of the synthesized images, and then applied the segmentation model to segment the tumors in the test set of the synthesized images. The results are shown in the right part of Table 5.5. Again, the segmentation model trained using our synthesized images shows better tumor segmentation performance than that trained using the images synthesized by the corresponding baseline methods of 3D cGAN and dEa-GAN. These experiments demonstrate the advantage of our sample-adaptive strategies, suggesting the characteristics of tumors may be better preserved in our synthesis.

5.3.6 Results on SISS2015

Table 5.6 reports the quantitative evaluation results on the SISS2015 dataset. As shown, the proposed dEa-SA-GAN performs best among the methods in comparison by considerable improvements that raises PSNR from 29.09 (Replica) to 31.08 (dEa-SA-GAN), lowers NMSE from 0.092 (Replica) to 0.061 (dEa-SA-GAN), and increases SSIM from 0.948 (Replica) to 0.961 (dEa-SA-GAN), respectively. The proposed SA-GAN method achieves the second-best performance. These experimental results demonstrate the superior synthesis quality of the images by the proposed methods on this dataset. Also, the proposed sample-adaptive strategy shows its effectiveness when comparing the proposed models with their corresponding baselines again. This is also verified by paired t-tests

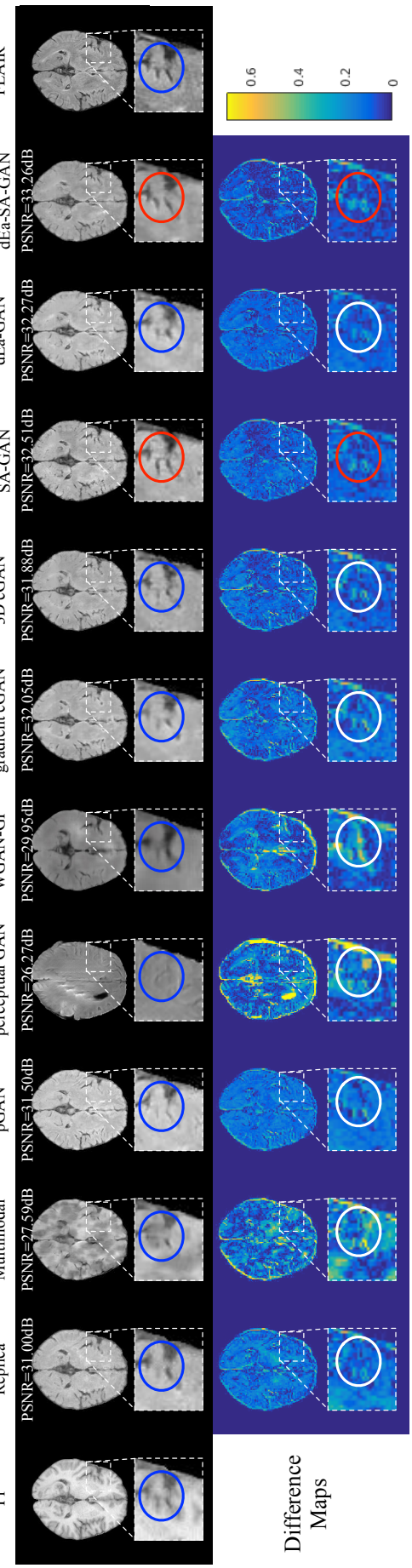


Figure 5.7: Comparison between the two proposed models and other state-of-the-art methods (T1 to FLAIR on SISS2015 dataset). The (a)-(j) regions in circles indicate where the visual difference between tissues is better perceived by SA-GAN and dEa-SA-GAN methods, as shown by the small values in the difference maps of (h) and (j).

Table 5.6: Quantitative evaluation results of the synthesized FLAIR-like images from T1 on the SISS2015 dataset (mean±STD). Paired t-tests are conducted between the proposed methods, i.e., SA-GAN and dEa-SA-GAN, and their baselines at the significance level of 0.05, respectively. When the improvement is statistically significant, the result of the proposed method will be underlined.

Methods	PSNR	NMSE	SSIM
Replica [41]	29.09±2.34	0.092±0.097	0.948±0.012
Multimodal [63]	29.45±2.21	0.094±0.110	0.948±0.013
pGAN [64]	29.59±1.83	0.090±0.090	0.947±0.011
perceptual GAN [32]	25.69±2.37	0.178±0.129	0.894±0.027
WGAN-GP [28]	29.42±1.60	0.092±0.107	0.949±0.011
gradient cGAN [148]	29.83±2.16	0.083±0.102	0.950±0.012
3D cGAN [148] (baseline 1)	29.82±2.15	0.084±0.110	0.950±0.012
SA-GAN (proposed)	**30.54±2.06**	**0.069±0.094**	**0.955±0.010**
dEa-GAN [148] (baseline 2)	30.41±2.16	0.073±0.089	0.956±0.011
dEa-SA-GAN (proposed)	**31.08±2.01**	**0.061±0.092**	**0.961±0.012**

at a significance level of 0.05 between the proposed models and their baselines. A visual example of the synthesized images is presented in Figure 5.7. Although all models generate FLAIR images with relatively good-quality, the visual difference between tissues in the synthesized images by SA-GAN and dEa-SA-GAN could be better perceived than those produced by the other state-of-the-art methods, as indicated by the circles in zoomed rectangles.

5.4 Discussion

In this chapter, we propose sample-adaptive GAN-based models aiming at improving the cross-modality MR image synthesis. In addition to the common global cross-modality mapping learned by existing GANs models, our proposed models have an additional sample-adaptive path to learn the local cross-modality mapping around a given sample. This allows our models to flexibly handle subtle sample-specific features for better synthesis. The proposed framework can be built upon various backbone networks to further improve them. The effectiveness of the proposed framework can be demonstrated by the advantages of SA-GAN and dEa-SA-GAN over their corresponding baselines. The improvement purely comes from the sample-adaptive path that provides auxiliary information to assist the synthesis of the baseline path (backbone network). Note that dEa-SA-GAN performs better than SA-GAN since the former builds upon a more powerful baseline dEa-GAN that utilizes edge information. When new and better backbone networks appear (which is not the focus of this chapter though), our sample-adaptive strategy could be integrated to further improve them. In this sense, our sample-adaptive models do not compete with but complement the existing GAN models. Moreover, from the ex-

perimental results, we can see that the advantages of our sample-adaptive GANs models become more salient in tumor regions than in the whole images. This shows the evident benefits of learning local sample space mapping to depict the useful fine details in the synthesis. Also, the proposed sample-adaptive models outperform the existing models in comparison, which include two GANs-based models. It is interesting to see that although the backbone network 3D cGAN performs worse than the gradient cGAN since 3D cGAN does not utilize gradient/edge information, its sample-adaptive counterpart SA-GAN outperforms gradient cGAN in both two datasets for all three synthesis tasks. This also reflects the benefits of our proposed sample-adaptive learning framework.

More discussions about the proposed framework are given as follows.

5.4.1 Discussion about Smoothness Assumption

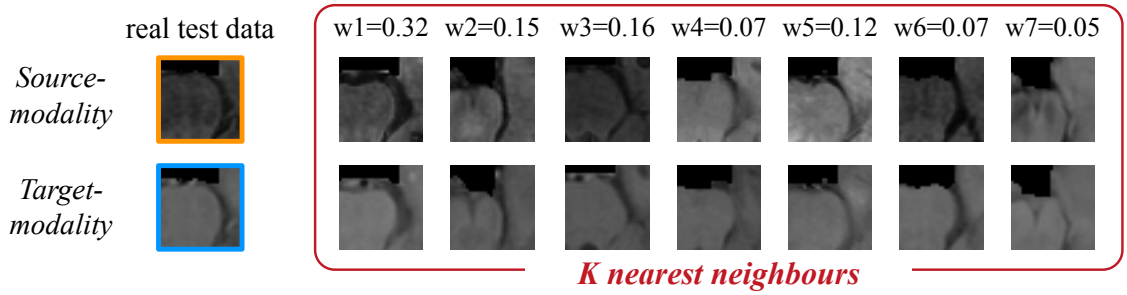


Figure 5.8: An example of test data, its nearest neighbors, and their generated weights.

In section 5.2.3, we assumed that the mapping from the source-modality space to the target-modality space is locally smooth. This smoothness assumption in sample space is a very common assumption used not only in the easier computer vision tasks, like classification, but also in the more difficult segmentation [149] and synthesis [40, 150]. This assumption also holds in medical image analysis, since medical images are quite similar in anatomy, and the shift from one subject to another is relatively small. Therefore, the classifiers for dense prediction (e.g, segmentation and synthesis) could be possibly interpolated from those of its neighboring samples effectively. More importantly, we only apply this smoothness assumption to interpolate the aligned corresponding image parts rather than the entire images. In addition, we provide an example of a test data from BRATS2015 for the T1-to-FLAIR synthesis task, its seven nearest neighbors, and their weights generated by our SA-GANs in Figure 5.8. It can be seen that the appearance of the found target-modality neighbors is very similar to that of the test data, which shows that the assumption works well in our SA-GANs.

5.4.2 Discussion about Labeled Samples

In our sample-adaptive path, the labeled samples (training samples) are used to find the KNNs of the input sample. The generality of the used labeled samples will affect the quality of the learnt local sample space mapping. When the labeled samples are very few and sparse in the sample space (e.g., a small number of samples far away from each other), we would not expect good performance from either the global or the local sample space mapping. But we would like to point out that, in the brain, there is relatively high redundancy in anatomical features of healthy tissue (as well as quasi-symmetry). Moreover, the images are often spatially co-registered (such as the images in BRATS2015) for analysis, which further makes them closer to each other and the interpolation of local image patches become possible. However, such information has not been fully utilized in the existing GANs models that learn a single global sample space mapping. When using a single global sample space mapping to capture both the transformation at the whole image level and the subtle and critical details around tumors, this mapping does become very complicated. As one of the motivations in this chapter, we decompose the learning of the global and the local sample space mappings to reduce the complexity of learning problem, with the baseline path learning the global sample space mapping as usual and the sample-adaptive path learning the local sample space mapping for further refinement. Unfortunately, there is no good a-priori rule on what number for samples would be need, and so careful validation is always required to ascertain the clinical utility of the technique and its limitations.

5.4.3 Difference between Non-local methods and Our Framework

The proposed sample-adaptive path in our framework is substantially different from non-local methods. The primary difference is that they work in two different spaces, i.e., the non-local methods work on the image pixels, while our method works in the image sample space. Although both methods learn weights for the local parts of feature maps, the non-local methods for static images focus on obtaining the long-range dependence among the pixels in one input image sample [151]. Differently, our SA-GANs are designed to capture the similarity between a training image sample and its candidate neighboring samples and generate weights to integrate the feature maps of the K nearest neighbors. In this way, we could capture the sample-specific features through the neighborhood-ship in the sample space.

5.5 Conclusion

This chapter points that a unified global model, which is trained for the whole-space mapping, could be insufficient for image synthesis due to the complexity of the prob-

lem, the scarcity of labeled data, and the variation among all the input image samples. To handle this issue, it proposes two novel end-to-end trained sample-adaptive GANs, i.e., SA-GAN and dEa-SA-GAN, for cross-modality lesion contained MR image synthesis. The proposed sample-adaptive learning strategy successfully extracts the particular features of each sample and learns its unique local sample space mapping for synthesis. Moreover, to build the better sample-specific path, the accessible real target-modality information is explicitly employed in both training and test stages during the learning of the local sample space mappings. The experimental results have demonstrated that our sample-adaptive framework can significantly improve the performance of the common GAN models and also outperform the state-of-the-art methods, including the commonly learnt GANs, in the recent literature on multiple MR image synthesis tasks.

Chapter 6

Learning Sample-adaptive Intensity Lookup Tables for Brain Tumor Segmentation

The effectiveness of additionally learning sample-adaptive mapping models for different samples has been demonstrated in Chapter 4. They could mitigate the problem that is brought from the whole sample-space mapping when a unified model is learnt to expect to match all the varied samples. It is worth exploring sample-adaptive learning for segmentation tasks. What if actively cope with the sample-specific visual variations among all the given images so that they can better adapt to the final segmentation task? For example, in the MR image segmentation task, the sample-adaptive intensity adjustment is learnt to mitigate the issue of too significant visual variations and also fit the adjusted samples to the subsequent segmentation. In this chapter, a CNNs based learning framework is proposed for brain tumor segmentation by MR images. It learns the sample-adaptive intensity lookup tables to dynamically transform the intensity contrast of input MR images, which could best support an optimal segmentation network and achieve the better performance.

6.0.1 Introduction

The location and appearance of brain tumors are key to the diagnosis and treatment of brain cancers. Such information is usually acquired by the non-invasive magnetic resonance imaging (MRI) and extracted by segmenting the tumor regions in the scanned images. However, accurate brain tumor segmentation is always challenging. For example, as one of the most aggressive brain tumors, glioma affects tens of thousands adults around the world [14, 152, 153]. Compared with some brain tumors like meningiomas, the extent of gliomas is more difficult to accurately define, due to their different shapes, sizes, and diffused locations which vary from patient to patient [6]. Moreover, gliomas of-

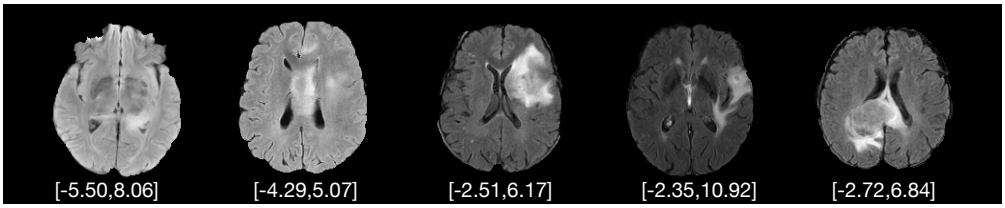


Figure 6.1: Tumor carried MR images of FLAIR modality preprocessed after the zero-mean and unit-STD normalization. Their normalized intensity scales are given in the bottom of images. After this linear rescaling, the significant intensity variation among MR images still remains.

ten extend their tentacle-like structures to invade the healthy brain tissues rather than just replacing them, which results in subtle changes and fuzzy tumor boundaries [11]. The accurate manual annotations of gliomas require laborious efforts from the professional radiologists [19] and could burden the workload in clinics. Therefore, automatic brain tumor segmentation methods on MR images would be beneficial in speeding up the process and provide objective and repeatable measurements for the radiation therapy treatment.

6.0.2 Motivation

The intensity values in MR images are not quantitative. Even the scanned intensities of the same tissue type can be distinctly different among MR images [154, 155]. In the brain tumor imaging, this results in the tumor-surround-contrast being much weaker in some MR images than others. This intensity variation inevitably increases the difficulty of training the segmentation model and generalizing it to any new MR images, which is the problem of interest in this work. To eliminate variations in image intensity, a preprocessing step, intensity normalization [155], is used to align the intensity values of MR images with a standard before image analysis is performed. This is also applied in some deep learning based approaches to preprocess the images for brain tumor segmentation [156]. Intensity normalization methods usually alter the histograms of MR images so that the discrepancy among these histograms is minimized after the normalization step [155, 157, 158]. However, this kind of intensity normalization approaches needs to be employed deliberately as an independent preprocessing step before segmentation. Since this step is not guided by the subsequent segmentation or interpretation task, some critical pathological clues in the original MR images may be adversely affected. Consequently, complicated intensity normalization steps are not commonly applied before the automatic MR image segmentation methods. Instead, a simple zero-mean and unit-STD (standard deviation) normalization is often adopted to preprocess MR images, which linearly rescales the intensity values. However, as shown in Figure 6.1, the intensity values still exhibit significant variation after this normalization.

This chapter proposes an end-to-end learning framework where sample-adaptive inten-

sity lookup tables (LuTs) are learnt to cope with differently contrasted MR images and promote brain tumor segmentation performance. The idea behind our work comes from our observation about how doctors deliver manual annotation. Relying on their expertise, the doctors usually adjust the intensity contrast of the whole MR image using software tools to highlight and make the regions of interest protruded, which helps better depict the contours of these regions. During this process, the doctors attempt to find a relatively ideal intensity mapping function through their naked (but trained) eyes for annotation. This procedure highly depends on the doctors' individual experience and the images are processed one by one. Inspired by this observation, we propose a learning framework that has two benefits. On one hand, it automatically learns the intensity-level mapping function (i.e., LuT) that suits the following segmentation task, which does not need the doctors' intervention; on the other hand, the learnt mapping function adapts to various MR images with different tumor-surrounding contrasts, therefore relaxing the training and generalization of the segmentation model to some extent. To be more specific, each individual MR image is particularly assigned with an intensity LuT, which is associated with a nonlinear mapping function that caters for the individual need of intensity adjustment. We explore two families of nonlinear mapping functions, i.e., piece-wise linear mapping functions and power mapping functions. The parameters of the mapping function are automatically predicted in our SA-LuT-Net, which vary with the input MR images. The intensity LuT and the subsequent segmentation model are trained together in an end-to-end manner only with the supervision of segmentation criterion. In this way, they can negotiate with each other to improve the segmentation performance. The superiority of our SA-LuT-Net is also demonstrated on both BRATS2018 and BRATS2019 [19] validation sets. The online evaluation results are used to validate that our method achieves better performance than many other state-of-the-art methods, while using fewer model parameters. Moreover, the transferring ability of the learnt LuTs is also evaluated by applying the LuTs learnt using one segmentation model to another segmentation model to improve the latter's performance. With this investigation, we could have an insight into the crucial visual adjustment knowledge that is learnt during in the proposed SA-LuT-Net.

6.1 Proposed method

6.1.1 Overview

This chapter proposes a SA-LuT-Net framework to explicitly handle MR intensity variation through learning sample-adaptive intensity LuTs for segmentation. An intensity LuT corresponds to a nonlinear mapping function that could be used to adjust the intensity levels of MR images from one set to another. In our case, a LuT takes the intensity levels from the given MR image as the input, and outputs the transformed intensity levels that

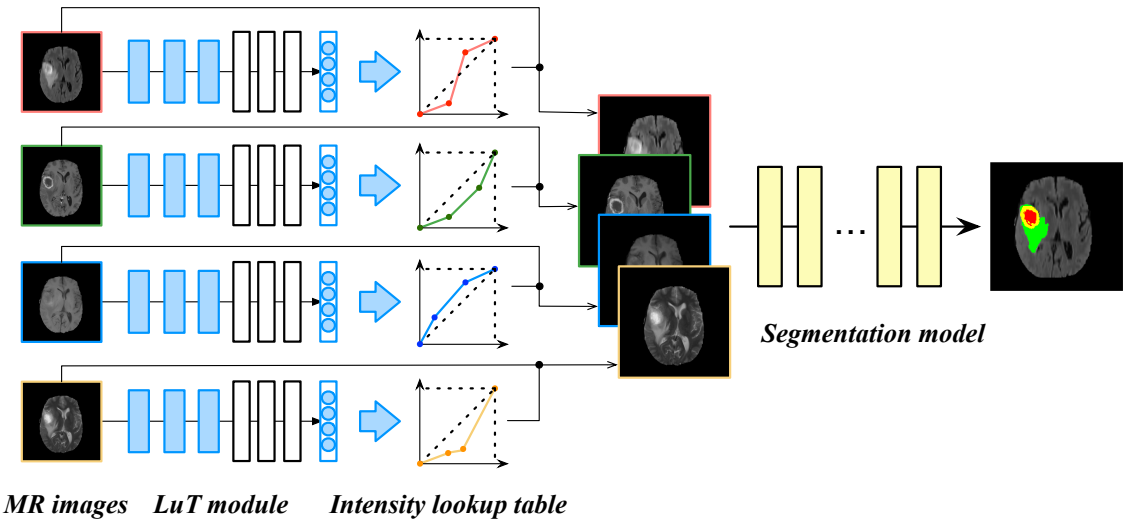


Figure 6.2: Overview of the proposed SA-LuT-Net framework under multi-modality scenario. It integrates two modules into the joint learning: (1) a LuT module generating the particular parameters of intensity mapping functions for every input MR image, and (2) a segmentation module processing the intensity adjusted MR images to estimate the labels of tumor regions. In this way, the learnt LuTs could help the input MR images become more suitable for the downstream segmentation task and improve the ultimate segmentation accuracy.

adjust the input contrast in the given MR image. In this way, the relevant visual information, like pathological tissues, could be better viewed. Due to the varying requirement of intensity adjustment from image to image, an input MR image may need a sample-specific mapping function and hence the corresponding LuT should be adaptive to different input image. In the proposed SA-LuT-Net, the learning of the sample-adaptive LuT is guided by the segmentation performance so that the learnt LuT could match the segmentation task. Figure 6.2 illustrates the overview of our proposed SA-LuT-Net framework. To be more specific, our framework incorporates a LuT module and a subsequent segmentation module. The LuT module generates the sample-adaptive parameters that determine different mapping functions of the LuTs. Each mapping function is then applied to the input MR image to change its intensity contrast. After that, the LuT-transformed image is further used as the input of the following segmentation module to predict the final segmentation labels. Through the end-to-end training in the proposed SA-LuT-Net, the LuT and the segmentation modules are jointly trained to negotiate with each other and achieve the optimal segmentation results.

6.1.2 Intensity LuT Module

Diverse nonlinear functions could be used to model the intensity transformation mappings of LuTs. Considering both the simplicity and flexibility, we choose two families of nonlinear functions, i.e., piece-wise linear functions and power functions, in this work.

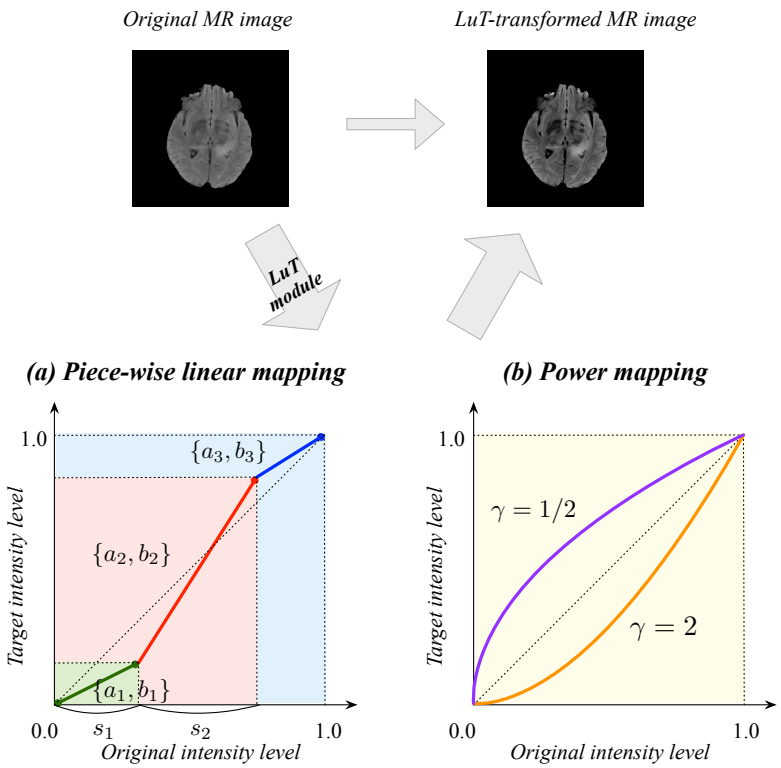


Figure 6.3: Two intensity lookup tables separately using (a) a piece-wise linear function and (b) a power function. Both of them can transform the input intensity levels in an MR image to the target levels by the estimated sample-specific parameters of mapping functions.

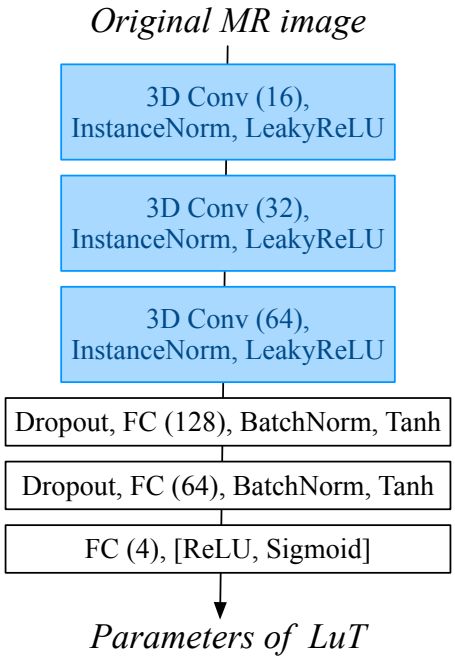


Figure 6.4: The architecture of a LuT module. It includes three convolutional blocks and also three FC blocks to predict the parameters of LuTs. In its last block, it uses different activation functions to constrain the value ranges of the parameters. Specifically, for the piece-wise linear functions, ReLu and Sigmoid layers are used to ensure the value ranges of the learnt parameters, while only RelU layer is applied for the power functions.

Piece-wise Linear Function

A three-segment piece-wise linear function is plotted in Figure 6.3 (a) as an example, and it could be mathematically formulated as:

$$\hat{\mathbf{x}} = \begin{cases} a_1\mathbf{x}+b_1, & 0\leq\mathbf{x}<s_1, \\ a_2\mathbf{x}+b_2, & s_1\leq\mathbf{x}\leq(s_1+s_2), \\ a_3\mathbf{x}+b_3, & (s_1+s_2)<\mathbf{x}\leq 1, \end{cases} \tag{6.1}$$

where the intensity levels in the given MR image and its LuT-transformed image are indicated by $\mathbf{x}$ and $\hat{\mathbf{x}}$, respectively. It is noted that, the real input intensity levels are linearly rescaled in each MR image, so that both $\mathbf{x}$ and $\hat{\mathbf{x}}$ are between $[0,1]$. The slope and the bias of the i-th line segment are denoted by a_i and b_i ($i=1,2,3$), respectively, while the horizontal intervals of the i-th line segment are s_i ($i=1,2$). In a special case, when $a_i=1$ and $b_i=0$, $\forall i$, the LuT becomes an identity mapping and the intensity levels of the given MR image will remain the same. The identity mapping function is illustrated as the dashed straight diagonal line in Figure 6.3 (a), indicating no intensity adjustment.

In the proposed SA-LuT-Net, the parameters of the mapping function are estimated by the LuT module. The LuT module processes a given input MR image and predicts the sample-specific values of the parameters a_i, b_i, and s_i for this input image. Taking the three-segment piece-wise linear mapping as an instance, our LuT module will output a four-dimensional vector, including a_1, a_3, s_1, and s_2 as its elements. In addition, according to the definition of mappings, the first line segment starts from the origin so that the parameters $b_1=0$, and other parameters a_2, b_2, and b_3 can be calculated from the output parameters of the LuT module. Combining the above together, Equation (6.1) can be rewritten as:

$$\hat{\mathbf{x}} = \begin{cases} a_1\mathbf{x}, & 0\leq\mathbf{x}<s_1, \\ \dfrac{a_3(s_1+s_2-1)-a_1s_1+1}{s_2}\mathbf{x} & \\ +\dfrac{(a_1-a_3)(s_1+s_2)s_1+(a_3-1)s_1}{s_2}, & s_1\leq\mathbf{x}\leq(s_1+s_2), \\ a_3\mathbf{x}-a_3+1, & (s_1+s_2)<\mathbf{x}\leq 1. \end{cases} \tag{6.2}$$

To estimate the parameters a_1, a_3, s_1, and s_2, our LuT module consists of three convolutional blocks and three fully connected (FC) blocks, as shown in Figure 6.4. Each convolutional block sequentially consists of a convolutional layer that applies kernels of size $4\times4\times4$ and the stride of four, an instance normalization layer, and also a LeakyReLU layer with the slope 0.2. These three convolutional blocks have 16, 32, and 64 output channels in order. In the first two FC blocks, a dropout layer using the rate of 0.5 is first employed, followed by an FC layer, a batch normalization layer, and a Tanh layer. The

outputs of these two FC blocks have 128 and 64 channels, respectively. For the last FC block, an FC layer that outputs 4 channels is applied, and different nonlinear activation functions are used to predict the four target parameters. To be specific, in order to build a regular one-to-one mapping function, we require $a_1 > 0$ and $a_3 > 0$, and s_1 and s_2 sit between zero and one according to definition. Therefore, the last FC block uses the ReLU function to guarantee the positive values of a_1 and a_3, and a Sigmoid function to ensure s_1 and s_2 in the range of $[0,1]$.

For the single-modality segmentation task, one LuT module using the three-segment piece-wise function possesses 0.57M trainable parameters. If using multi-modality MR images, each modality will be associated with a LuT module. This design considers that separately using different LuT modules for each input modality helps to focus on the specific contrast characteristics of each modality for the segmentation. Therefore, our learnt LuT adapts to different samples and different modalities.

Power Function

Power functions are also investigated as the nonlinear intensity mapping function associated with LuTs, which are as follows and illustrated in Figure 6.3 (b).

$$\hat{\mathbf{x}} = \mathbf{x}^{\gamma}, \quad \text{where } 0 \leq \mathbf{x} \leq 1, \tag{6.3}$$

where the power γ is the only learnable parameter that controls the LuT curve for each input MR image. The curve of the power function passes through the two points (0,0) and (1,1), and here we require $\gamma > 0$. When $\gamma = 1$, there will be no intensity adjustment, shown as the dashed diagonal line in Figure 6.3 (b). When $\gamma > 1$, the power function has a convex shape for $\mathbf{x} \in [0,1]$; when $0 < \gamma < 1$, the power function has a concave shape for $\mathbf{x} \in [0,1]$. Figure 6.3 (b) also gives two LuT curves of $\gamma = 2$ and $\gamma = 1/2$ as examples. Compared with the piece-wise linear functions mentioned above, these power functions have fewer parameters to learn, but may also have less flexibility for the adjustment. Power functions are also used in gamma correction, as a nonlinear operator to correct the brightness/intensity values of the pixels in natural image photographing and the display of medical images [159–163].

For power functions, we also use the architecture of the LuT module in Figure 6.4 to predict the parameter γ, except that the number of the output channel in the last layer of LuT module becomes one. As γ is defined to be larger than zero, the ReLU function is applied to enforce it to be a non-negative value for the LuT transformation.

6.1.3 Segmentation Module

In the right part of Figure 6.2, our segmentation module is illustrated. Compared with those stand-alone segmentation models, ours takes the intensity-adjusted MR images as its input and is jointly trained with the LuT module. Here, our SA-LuT-Net framework is developed upon two backbone networks, respectively, which are two state-of-the-art models for brain tumor segmentation. Both of them have been reviewed in Chapter 2 in this thesis. The first one is the modified 3D Unet model [17], which contains 26 convolutional layers. It incorporates residual connections in its encoding convolutional blocks to combine the neighboring shallow and deep features, and also integrates different-depth segmentation outputs in an element-wise summation way to apply deep supervision for the final model output. The modified 3D Unet achieved the third place in BRATS2017 [19]. The second backbone network is DMFNet [18]. It has 69 convolutional layers built as a 3D Unet-like structure with skip connections. It replaces the ordinary convolutional layers in its first six encoding residual units with more efficient adaptive dilated multi-fiber layers to capture the multi-scale feature representations from brain tumor images. It attained the comparable results on the BRATS2018 validation set [19] with the challenge first-place model, NVDLMED [113], but only includes about 1/10 learning parameters of NVDLMED. As for the detailed architectures of the modified 3D Unet and DMFNet, please be referred to the original papers [17] and [18].

6.1.4 Training Strategy

Similar to the common segmentation models, the proposed SA-LuT-Net only exploits the ultimate segmentation loss to guide its learning process. Since the LuT module is in the beginning layers (close to the input) of the entire model and the following segmentation module is usually very deep, the training of our SA-LuT-Net may encounter the gradient vanishing issue. That is, the loss gradient is not easily propagated properly to the LuT module. To alleviate this issue, we implement a three-stage training strategy. At the first stage, the segmentation module is trained as a stand-alone model without the LuT module. At the second stage, an alternative and iterative training approach is applied on the LuT and segmentation modules. To be more specific, in one epoch the LuT module is updated while the segmentation module is fixed, and then in the next epoch, the LuT module is fixed and only the segmentation module is updated. This will be iterated for a number of epochs to gradually train both of the LuT and the segmentation modules. Since the segmentation module is frozen during the updating of LuT module, the gradient vanishing issue in the LuT module will be mitigated. In addition, as the segmentation module is much deeper than a LuT module, it requires more rounds of updating before convergence. Therefore, after the LuT module is fully trained and becomes stable in the second stage, it will be set frozen at the third stage and only the segmentation module

keeps updating for a few more rounds. After all these three stages, our SA-LuT-Net completes its training process.

6.1.5 Remarks

It is noteworthy that the proposed sample-adaptive LuT framework is essentially different from the attention mechanism used in some image segmentation models [164–166]. First, they are applied in different spaces. The attention-based methods weigh the importance of different spatial regions in the convolutional feature maps, while our method highlights the segmentation-related intensity contrast in the intensity space. For example, attention commonly assigns the similar weights to spatially neighboring pixels. In contrast, our LuT transforms the pixels of a given intensity level in the same manner, even if they are spatially distant. Second, attention is usually integrated into the deep layers of the CNNs after the convolutional feature representations have been extracted for the segmentation task. In contrast, our sample-adaptive LuT is applied at the very early layers of the deep model to preserve the critical visual clues from the beginning. Moreover, attention and our sample-adaptive LuT do not compete with each other, instead they could be applied together to promote segmentation performance from different perspectives.

6.2 Experimental Results

6.2.1 Dataset and Training Settings

We evaluate the proposed SA-LuT-Nets framework on BRATS2018 and BRATS2019 [19] datasets. The former consists of a training set with 285 subjects and a validation set with 66 subjects, and the later contains 335 subjects in its training set and 125 subjects in its validation set. Each subject has four-modality $240 \times 240 \times 155$ MR images, i.e., T1, FLAIR, T2, and post-contrast T1-weighted (T1ce). For the training sets, every subject also contains its corresponding segmentation ground truth including background, necrotic and non-enhancing tumor, peritumoral edema, and GD-enhancing tumor, whose labels are 0, 1, 2, and 4, respectively. To compare the segmentation results with the ground truth, the whole tumor (WT) with labels 1, 2 and 4, the tumor core (TC) with labels 1 and 4, and the enhancing tumor (ET) with label 1 are used as three different tumor regions. For the validation sets, since the segmentation ground truth is not given, its results need to be evaluated on BRATS2018 and BRATS2019 online server.

We separately apply the preprocessing steps of the two backbone models on the input MR images and then use the proposed SA-LuT-Nets for the LuT transformation and the final segmentation. Only the original segmentation losses of modified 3D Unet and DMFNet are used to separately train these two backbones based SA-LuT-Nets. The learn-

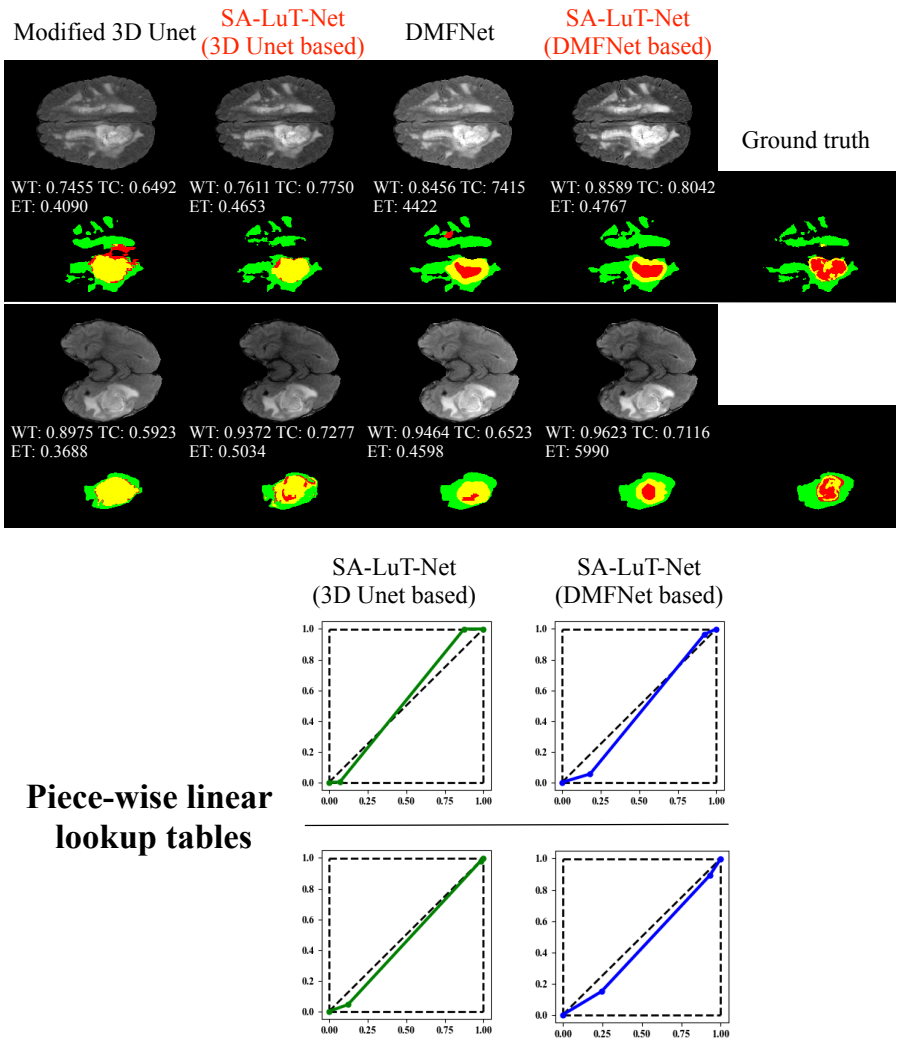


Figure 6.5: Comparisons between SA-LuT-Nets using piece-wise linear mapping functions (three line segments) for LuTs and the baselines.The displayed images in first and third rows are preprocessed for the baselines and preprocessed and LuT-transformed for the SA-LuT-Nets, respectively. The second and fourth rows give their corresponding segmented labels. These learnt LuTs are flexible to adaptively adjust the intensity levels of the FLAIR MR images for the brain tumor segmentation task.

ing rates of the LuT module and the segmentation module are fixed as 0.001 and 0.0001, respectively, through the Adam optimizer in our framework. The second and third training stages separately have 150 and 250 training epochs. No any new hyper-parameter is involved in our SA-LuT-Nets except from those used in the backbone models.

6.2.2 Comparison with Baselines

We study the effectiveness of the proposed SA-LuT-Nets framework with both the piece-wise linear mapping function and the power mapping function, and compare them with the two baseline models, i.e., the modified 3D Unet [17] and DMFNet [18], respectively. For the piece-wise linear function, we employ a three-segment piece-wise linear function in

Table 6.1: Dice scores of FLAIR segmentation results on BRATS2018 training set, reported by mean(std).

Methods	WT	TC	ET
Modified 3D Unet [17] (baseline)	0.8437(0.1520)	0.5912(0.2104)	0.3520(0.2342)
SA-LuT-Net (power function, 3D Unet based, ours)	0.8512(0.1093)	0.6372(0.1766)	0.3869(0.2554)
SA-LuT-Net (piece-wise linear function, 3D Unet based, ours)	**0.8621(0.1258)**	**0.6450(0.1968)**	**0.3959(0.2647)**
DMFNet [18] (baseline)	0.8549(0.1031)	0.5499(0.2408)	0.3696(0.3055)
SA-LuT-Net (power function, DMFNet based, ours)	0.8601(0.1168)	0.5914(0.2419)	0.3659(0.2806)
SA-LuT-Net (piece-wise linear function, DMFNet based, ours)	**0.8746(0.0864)**	**0.6459(0.2353)**	**0.3776(0.2944)**

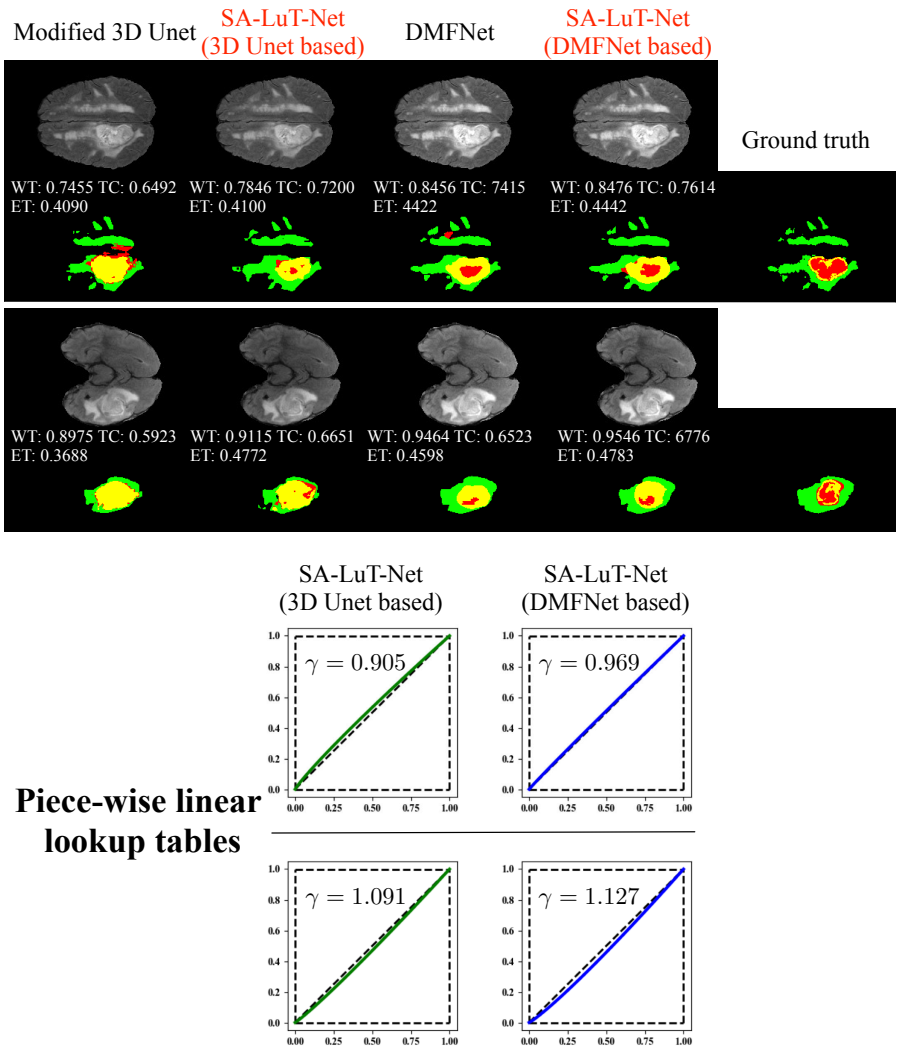


Figure 6.6: Comparisons between SA-LuT-Nets using power functions for LuTs and the baselines. The displayed images in first and third rows are preprocessed for the baselines and preprocessed and LuT-transformed for the SA-LuT-Nets, respectively. The second and fourth rows give their corresponding segmented labels. The curves of the learnt power mapping LuTs are different according to the input MR images, making these images more suitable for brain tumor segmentation.

this experiment. In this comparison, the experiments are conducted for a single modality FLAIR segmentation task. For evaluation, the BRATS2018 training set is separated as five folds to cross validate. Table 6.1 reports the Dice score results of these compared methods on three tumor regions. The higher values of dice scores indicate better segmentation results.

As can be seen, our SA-LuT-Nets with both LuT mapping functions achieve overall significant improvements over their corresponding baselines. Especially, SA-LuT-Net (piece-wise linear function) based on 3D Unet increases 1.8% on WT, 5.4% on TC, and 4.4% on ET from 3D Unet, meanwhile SA-LuT-Net (piece-wise linear function) based on DMFNet improves 2.0% on WT, 9.6% on TC, and 0.8% on ET from DMFNet. We can also find that the improvement on segmenting TC region is the most salient, by learn-

Table 6.2: Comparisons between SA-LuT-Net (DMFNet based) and intensity normalization approach. The Dice scores of FLAIR segmentation results are reported by mean(std).

Methods	WT	TC	ET
DMFNet [18]	0.8549(0.1031)	0.5499(0.2408)	0.3696(0.3055)
DMFNet+Norm [155]	0.8494(0.0979)	0.5204(0.3110)	0.3769(0.2909)
SA-LuT-Net (DMFNet based)	**0.8746(0.0864)**	**0.6459(0.2353)**	**0.3776(0.2944)**

ing the proposed sample-adaptive LuTs based on both two backbone models. Further comparing the results obtained by different LuT mapping functions, we can see that the three-segment piece-wise linear mapping function performs better than the power function. Compared with the power function that has only one parameter, the three-segment linear function has four learnt parameters, which has more flexibility for intensity adjustment and therefore could better negotiate with the segmentation module to improve segmentation results. Two visual examples with these two mapping functions are separately provided in Figures 6.5 and 6.6, showing the learnt LuTs adaptively varing with the input MR images and enhancing their tumor regions to help the segmentation task.

6.2.3 Comparison with Intensity Normalization

To study the effectiveness of our end-to-end learnt LuTs, we compare the proposed SA-LuT-Net (DMFNet based) using the three-segment piece-wise linear function with the traditional intensity normalization approach [155] as the preprocessing of MR images for the FLAIR segmentation task via the five-fold cross-validation experiments on BRATS2018 training set. Their results are reported in Table 6.2. As shown, the segmentation performance is not improved by standardizing the input MR images using the normalization approach [155]. This may be caused by the fact that the intensity normalization is conducted as a standalone step to segmentation. Hence, some pathological details that are critical to the following segmentation may be altered and ignored. In contrast, our end-to-end learnt LuTs based method can increase the Dice scores, which validates the effectiveness of the communication between the LuT module and the subsequent segmentation module. Because of this communication, the parameters of LuTs could be adjusted with the guide of segmentation task and help the LuT-transformed images suit the final segmentation. Two visual comparisons are shown in Figure 6.7. We can see that the traditional intensity normalization [155] tends to standardize the MR images, while our intensity LuTs enhance the contrast between different tissues according to the input MR images and highlight the tumor information for the segmentation. Both the quantitative and the visual results demonstrate the advantages of our proposed end-to-end trained SA-LuT-Net in segmentation tasks.

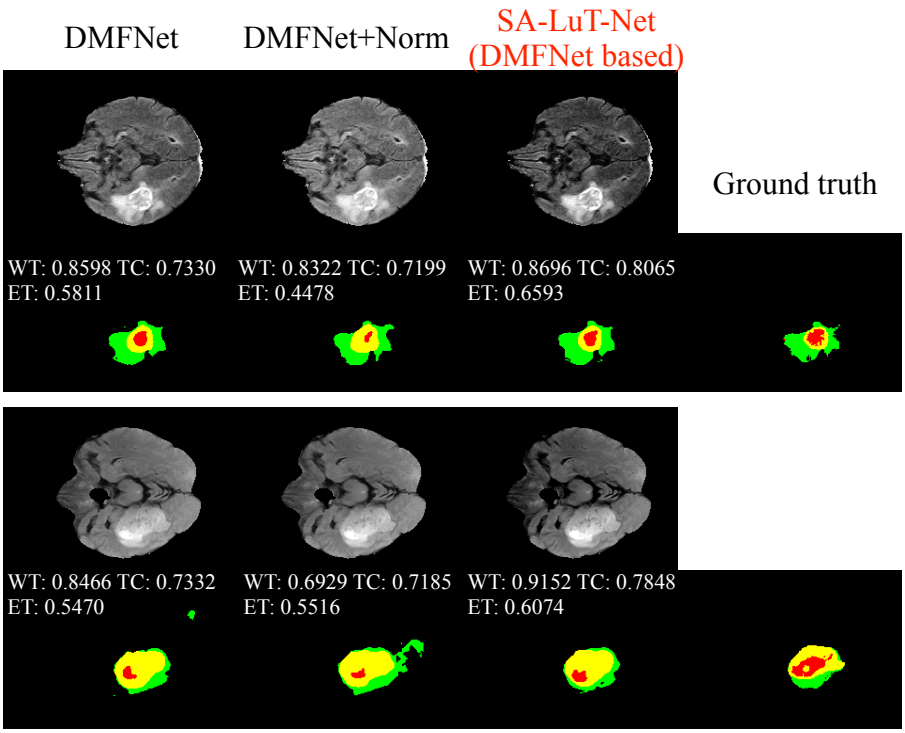


Figure 6.7: Comparisons between SA-LuT-Net (DMFNet based) and intensity normalization approach [155]. The displayed images in first and third rows are preprocessed for the baseline and preprocessed and LuT-transformed for the SA-LuT-Net. For the intensity normalization used approach, they are standardized and preprocessed. Using the proposed end-to-end learnt LuTs, the LuT-transformed MR images have more protruded tumor regions and get better segmentation results.

6.2.4 Comparison with the State-of-the-arts

Table 6.3: Multi-modality tumor segmentation results on BRATS2018 validation set.

Methods	Dice scores			Hausdorff95		
	WT	TC	ET	WT	TC	ET
S3D-UNet [114]	0.8935	0.8309	0.7493	-	-	-
Kao et al. [167]	0.9047	0.8135	0.7875	4.32	7.56	3.81
DMFNet [18] (baseline)	0.9062	0.8454	0.8012	4.66	6.44	3.06
No New-Net [112]	0.9083	0.8544	0.8101	4.27	6.52	**2.41**
NVDLMED [113]	0.9068	0.8602	**0.8173**	4.52	6.85	3.82
SA-LuT-Net (DMFNet based)	**0.9116**	**0.8746**	0.8073	**3.84**	**5.16**	3.67

To widely compare the proposed SA-LuT-Net framework with the existing successful brain tumor segmentation models, the experiments of four-modality tumor segmentation task are also separately applied on the BRATS2018 and BRATS2019 validation sets. We use the SA-LuT-Net (DMFNet based) with the three-segment piece-wise linear mapping in this investigation, considering that on this task DMFNet performed significantly better than the modified 3D Unet and the piece-wise linear function is more flexible than the power function. Two different segmentation evaluation metrics, i.e., Dice score and

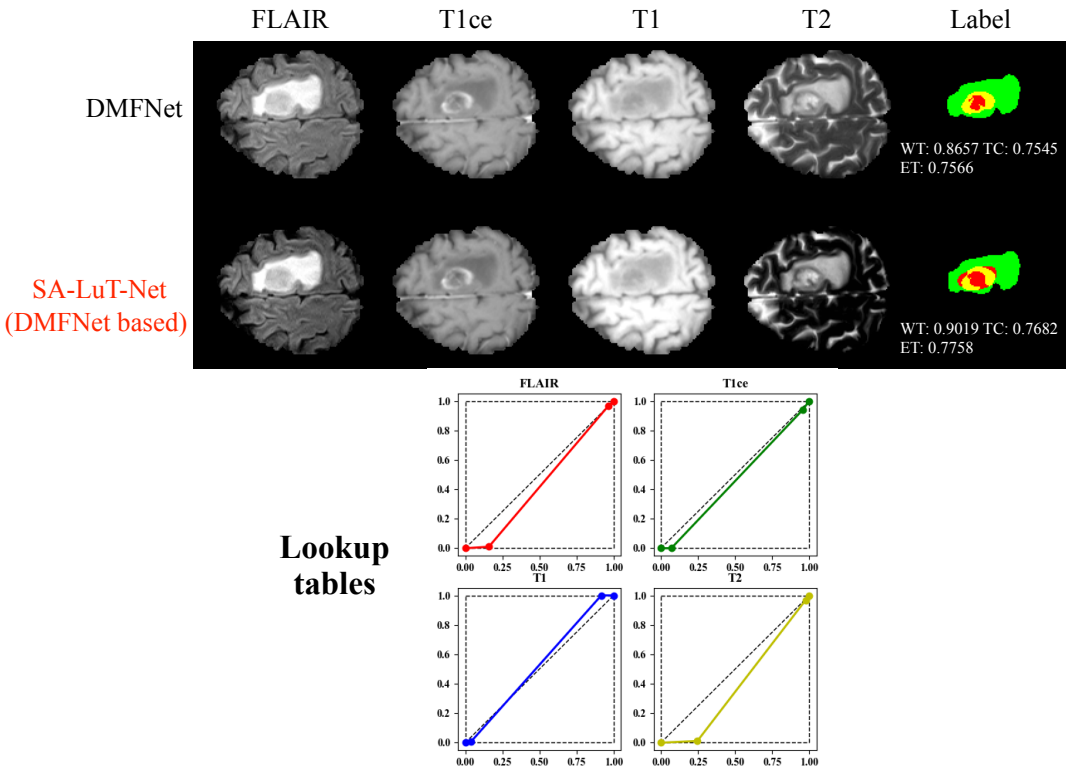


Figure 6.8: A qualitative example from BRATS2018 validation set. The displayed images are preprocessed for the DMFNet and preprocessed and LuT-transformed for the proposed SA-LuT-Net. The LuT curves are learnt differently for the four-modality images. The tumor tissues are more protruded, and more easily recognized by the segmentation network after the LuT transformation in the proposed SA-LuT-Net framework.

Hausdorff95 distance, are measured and reported by BRATS2018 and BRATS2019 online server, as other methods in comparison do. The lower value of the Hausdorff95 distance means better segmentation result. This work focuses on comparing the proposed SA-LuT-Net with the state-of-the-arts single-model based approaches on BRATS2018 dataset. Table 6.3 presents their results. It can be seen that our SA-LuT-Net (DMFNet based) performs best among all the compared models in the segmentation of WT and TC tumor regions in terms of the two evaluation metrics. It again achieves better segmentation results than its baseline DMFNet [18] on all the three regions. The improvement is especially significant on TC tumor region with 2.9% Dice score. The proposed SA-LuT-Net only attains slightly inferior performance on ET region to NVDLMED [113] and No New-Net [112], which are the first- and second-place winners in BRATS2018 challenge. Whereas, the proposed SA-LuT-Net has much fewer parameters (6.14M) to learn. This is in contrast to NVDLMED and No New-Net, which have 40.06M and 10.36M parameters, respectively, as calculated in [18]. Overall, our SA-LuT-Net achieves better performance than the state-of-the-art single models and uses relatively fewer learning parameters. For BRATS2019 validation set, since the aforementioned models did not formally present their results and the first- and third-place winners of BRATS2019 use

Table 6.4: Multi-modality tumor segmentation results on BRATS2019 validation set.

Methods	Dice scores			Hausdorff95		
	WT	TC	ET	WT	TC	ET
Wang et al. [168]	0.8940	0.8070	0.7370	5.68	7.36	5.99
Li et al. [169]	0.8860	0.8130	0.7710	6.23	7.41	6.03
Xue et al. [170]	0.9000	0.8300	0.7500	6.13	6.77	5.07
Zhao et al. [171]	**0.9100**	0.8350	0.7540	4.57	5.58	3.84
DMFNet [18] (baseline)	0.9000	0.8018	0.7706	5.22	7.52	**3.30**
SA-LuT-Net (DMFNet based)	**0.9079**	**0.8482**	**0.7821**	**4.46**	**5.26**	3.69

Table 6.5: The effect of LuT line-segment numbers using SA-LuT-Net (DMFNet based) on FLAIR segmentation.

Methods	WT	TC	ET
Baseline	0.8602	0.5219	0.4251
Two line segments	0.8594	0.5323	0.4095
Three line segments	**0.8856**	**0.6494**	**0.4270**
Four line segments	0.8825	0.6213	0.4196

ensemble models, we compare the proposed SA-LuT-Net based on DMFNet with four well-performed single models including the second-place winner [171] in the challenge. Their results are quoted in Table 6.4. As can be seen, although the model from [171] performs slightly better than our SA-LuT-Net on WT region, the proposed model largely outperforms it by 1.3% on TC and 2.8% on ET regions. Thus, the proposed SA-LuT-Net (DMFNet based) shows the overall consistent superiority over the compared models on BRATS2019 dataset. Figure 6.8 gives a visual example from the BRATS validation set, where the LuTs are learnt differently varying with the four MRI modalities. With the transformation by the learnt LuTs, the tumor tissues are more contrasted, and are better discriminated by the subsequent segmentation network.

6.2.5 Ablation Studies

Study about the Number of Line Segments

An ablation study is conducted to investigate how the number of line segments used in piece-wise linear mapping LuTs affects the segmentation performance. In this experiment, the single modality FLAIR segmentation task is used with the training subjects of BRATS2018 randomly separated into 80% training data and 20% test data. Two-segment, three-segment and four-segment piece-wise linear functions are tested in the proposed SA-LuT-Net (DMFNet based). Their results are compared in Table 6.5. As reported, using the same training setup, the model with the three-line-segment LuTs achieves the best results. Meanwhile, the two-line-segment LuT gets similar results as the baseline but

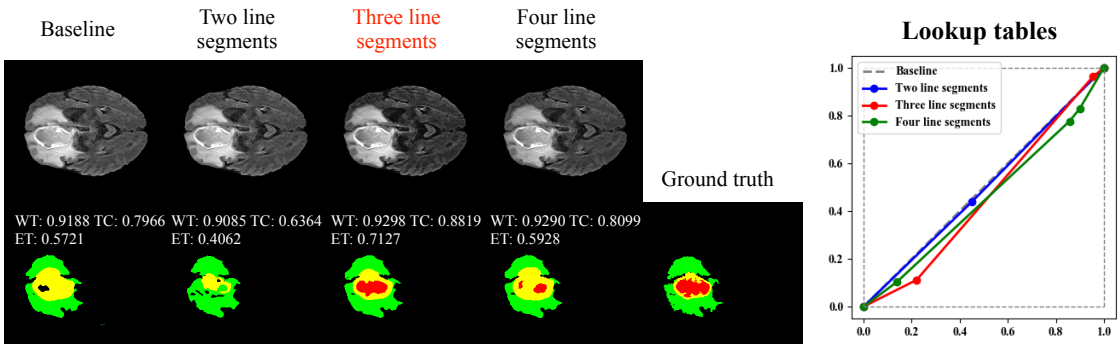


Figure 6.9: A visual example of using different segment numbers. All the three curve lines show a concave shape for this MR sample, and using the three-line segments for LuTs gets the best segmentation results.

Table 6.6: The effect of different learning constraints in three-line-segment LuT using SA-LuT-Net (DMFNet based) on FLAIR segmentation.

Methods	WT	TC	ET
Baseline	0.8602	0.5219	0.4251
Single linear line	0.8425	0.5289	0.3669
Intensity clipping	0.8664	0.5339	0.4179
Three line segments	**0.8856**	**0.6494**	**0.4270**

it is inferior to the LuTs employing three or four line segments. This is possibly caused by the less flexibility of the two-segment piece-wise linear mapping function that cannot sufficiently model a sought-after non-linear intensity transformation. Besides, a visual example of the results using different numbers of line segments is provided in Figure 6.9. All the three curve lines show a similar shape for this MR sample, and using the three-line segments for LuTs gets the best segmentation results.

Study about the Learning Constraints of Line Segments

We also conduct another ablation study to investigate the effect of using different constraints during learning LuTs on segmentation results. In this experiment, we also apply randomly selected 80% data from BRATS2018 for training and the rest 20% for test on the FLAIR segmentation task. We compare the proposed three line segments with the single linear line and the intensity clipping in LuTs. The single linear line setting means learning a linear mapping of the intensity levels from $[0,1]$ to $[0,z]$. Here, $z>0$ is learnt by the LuT module for every image. The intensity clipping setting indicates a three-piece-wise linear function but limiting $a_1=0$ and $a_3=0$. Their results are presented in Table 6.6. As can be seen, using the single linear line setting, the WT segmentation Dice score even drops by 4.3% from the baseline. After looking into the detailed results of every image, we find that the learnt sample-adaptive z varies largely among samples. This may be caused by two reasons. First, the intensities in $[0,z]$ could be scaled to $[0,1]$ by dividing

z. This changes the scale of the absolute values, but the relative tissue contrast remains. In practice, we care more about the latter for segmentation. Another reason is when z is not constrained, the optimization could easily go unbounded. Also, the results show that applying the intensity clipping setting, the learnt LuTs can slightly promote DMFNet by 0.6% on WT, but the improvement is largely lower than that by the setting without clipping (2.5%). This is possibly due to the less flexibility with clipping. Therefore, the proposed three-piece-wise learning constraint is a more suitable choice in this case.

6.2.6 Study about Transferring LuTs between Segmentation Models

Moreover, we are interested in investigating whether the learnt LuTs could be general, to some extent, to the segmentation task. Therefore, we conduct an experiment to test if the LuTs learnt using one segmentation model could be used to improve the performance of another segmentation model. Specifically, we first obtain the LuTs (using the three-segment piece-wise linear function) learnt by the modified 3D Unet/DMFNet, apply them to transforming the MR images, and directly use these transformed images to train the other segmentation model DMFNet/modified 3D Unet correspondingly for brain tumor segmentation using the single imaging modality FLAIR. The Dice scores of their segmentation results via five-fold cross validation are reported in the last two rows of Table 6.1. When comparing them with the results in first four rows, we can find that this transferring approach improves the segmentation performance on WT and TC regions from the baselines, 3D Unet (1.0% for WT and 3.6% for TC) and DMFNet (1.3% for WT and 6.1% for TC). As for the ET region, this approach gets worse or similar results compared with baselines, indicating that the transferring LuTs without end-to-end training with the segmentation model may not be sufficient to depict the required contrast to segment the very small region of ET. In addition, the results of the proposed SA-LuT-Nets are better than those of the transferring approach on all the three tumor regions. This further verifies the positive effect of the end-to-end training in the proposed SA-LuT-Net framework.

6.3 Discussion

In this chapter, we propose to learn sample-adaptive LuTs that dynamically adjust the intensity contrast of MR images and improve the brain tumor segmentation performance. The proposed SA-LuT-Net framework complements the novel LuT module that learns the sample-specific parameters of the nonlinear intensity mapping with a segmentation module to segment on the LuT-transformed images for the final task. The entire framework is trained in an end-to-end manner which allows the learnt LuTs to flexibly transform the MRI contrast for better segmentation. The proposed framework can be developed on

Table 6.7: FLAIR segmentation results of transferring LuTs between segmentation models, reported by mean(std).

Methods	WT	TC	ET
Modified 3D Unet [17] (baseline)	0.8437(0.1520)	0.5912(0.2104)	0.3520(0.2342)
SA-LuT-Net (3D Unet based, ours)	0.8621(0.1258)	0.6450(0.1968)	0.3959(0.2647)
DMFNet [18] (baseline)	0.8549(0.1031)	0.5499(0.2408)	0.3696(0.3055)
SA-LuT-Net (DMFNet based, ours)	0.8746(0.0864)	0.6459(0.2353)	0.3776(0.2944)
Transferring LuTs from 3D Unet to DMFNet	0.8682(0.1339)	0.6109(0.2676)	0.3302(0.2930)
Transferring LuTs from DMFNet to 3D Unet	0.8537(0.1474)	0.6272(0.2059)	0.3572(0.2681)

various segmentation backbone networks to further improve them. Its effectiveness was verified by the performance superiority of the proposed SA-LuT-Nets over the baselines, i.e., the modified 3D Unet and DMFNet. The improvement purely comes from the learnt LuT transformation that enhances the task-related contrast to support the segmentation module (backbone network). The learnt LuTs that are acquired from the training with one segmentation network can be transferred to apply to another segmentation network, showing their relatively consistent improvements on different segmentation networks. Also, for multi-modality MR images, the proposed SA-LuT-Net generates LuTs that vary with each modality and each sample to achieve the freedom for intensity adjustment. The four-modality segmentation results by the proposed SA-LuT-Net (DMFNet based) are overall better than other state-of-the-art models with different architectures, although the comparison with segmentation models of various structures is not the focus of this chapter. It is anticipated that when more advanced brain tumor segmentation models appear, the proposed SA-LuT-Net framework could be built upon them and further improve them. In addition, the reason why the proposed three-segment piece-wise linear mapping function performs better than the power function may be similar to the reason why it outperforms the piece-wise linear function of two line segments. The power function has only one learnable parameter and the two-segment piece-wise function has two learnable parameters, which have less freedom to adjust the intensity levels for segmentation, compared with the recommended three-segment piece-wise linear function. The four-segment piece-wise linear function gets similar results as the three-segment one, showing that the LuTs using four learnt parameters may be sufficiently flexible on this brain tumor segmentation dataset. As mentioned, the used piece-wise linear mapping function and power function are not the only choices to transform the intensity levels. More functions depicting higher degree of nonlinearity could be explored in the proposed SA-LuT-Net framework. In addition, from the experimental results, we can see that the advantage of our SA-LuT-Net is particularly salient on the tumor core part over almost all the compared models. This shows the evident benefits of the end-to-end learnt LuTs that help to highlight the essential tumor core details, while some of these details in the original MR images may not be well viewed due to relatively poor contrast.

6.4 Conclusion

This chapter proposes a novel sample-adaptive learning framework, i.e., SA-LuT-Net, for brain tumor segmentation. It learns the optimal sample-adaptive intensity LuTs to actively mitigate the significant visual variations among the different input MR images and dynamically adjust their intensity contrasts according to their input values to match with the subsequent segmentation network. The ultimate aim of the proposed framework is to increase the MR image segmentation performance through the sample-adaptive learning

strategy. The proposed framework is separately implemented upon two prevalent backbone segmentation networks. The experimental results on two public datasets demonstrate that the proposed SA-LuT-Nets successfully improve the performance of the common segmentation models (backbones) and also outperform the state-of-the-art models from the recent literature for brain tumor segmentation.

Chapter 7

Conclusions and Future Work

In this chapter, the main contributions of this thesis will be concluded. Also, the potential research directions of its future work will be discussed.

7.1 Conclusion

This thesis focuses on two per-voxel prediction tasks, i.e., image synthesis and segmentation, on medical images by deep convolutional neural networks (CNNs). Targeting at these two challenging tasks, adversarial learning and sample-adaptive learning frameworks are explored in the thesis to promote the prediction performance upon the deep CNNs from the following three perspectives: (1) designing effective deep 3D CNNs based GAN models to learn the volumetric medical image mapping for the per-voxel regression task (cross-modality MR image synthesis); (2) exploring more advanced GANs to preserve the vital brain structural details in the synthesized image for better per-voxel regression (cross-modality MR image synthesis) performance; (3) developing a GANs based sample-adaptive learning framework to learn a specific model for each image sample for per-voxel regression (lesion contained cross-modality MR image synthesis); (4) establishing a novel deep CNNs based learning framework to sample-adaptively mitigate the significant visual variation among MR images for the challenging per-voxel classification task (brain tumor segmentation on MR images). To be more specific, this section will conclude the proposed works in detail as follows.

- In Chapter 3, adversarial learning is investigated in the designed 3D CNNs based GAN model to learn the mapping for cross-modality brain MR image synthesis. The study about the 2D and 3D CNN structures in GAN models provides a solid guidance of designing per-voxel regression models for MR images. The take-away message is that using the 3D architecture can mitigate the discontinuous estimation across the 2D slices in 3D medical images and promote their voxel-wise prediction performance.

- In Chapter 4, beyond the 3D CNNs based GANs from Chapter 3, more advanced edge-aware adversarial learning strategies are proposed for cross-modality brain MR image synthesis. The effectiveness and importance of preserving the edge information show that only minimizing the voxel-wise intensity similarly is not sufficient to learn a well-performed synthesis mapping. Therefore, two adversarial learning strategies of integrating the edge maps into the GANs can enforce the synthesized images much sharper to reflect more brain structural details. In addition, our investigation also finds that incorporating the edge information into the adversarial learning of both generator and discriminator better serves the synthesis. Its effectiveness is demonstrated on the two different brain MR image datasets. Last but not the least, the generality of the designed edge-aware GANs' 2D variants is verified across different generic image datasets. Thus, object contours can be sufficiently captured during the adversarial learning of CNNs based GANs in the synthesis of either medical or generic images.

- In Chapter 5, the weakness of training a unified model for all the input image samples with high variations is pointed out. Based on this, a novel GANs based sample-adaptive learning framework is developed. It actively learns the specific characteristic of each sample via its unique local sample-space mapping, by utilizing the relationship between the input sample and its neighboring training samples. This sample-adaptive learning is built on top of the common whole sample-space learning to also exploit the common features of samples. The effectiveness of the proposed learning framework is demonstrated on two lesion contained MR image datasets, which validates that the framework successfully copes with the learning issue in the existing whole sample-space mapping based GANs for per-voxel synthesis. This thesis investigates not only the quality of finally synthesized images but also the intermediate results to study the implicit knowledge learnt by the proposed local sample-space mapping and analyze the reasons of its effectiveness. Besides, this thesis points another learning issue in the most existing GANs. They only use the target-modality information of the training samples in error evaluation during training but not actively exploit this crucial ground-truth information to directly serve the synthesis. In contrast, the proposed framework utilizes the real target-modality training samples not only to evaluate loss functions but also to learn target-modality-related features to help synthesis. In addition, the performance of the synthesized lesion in the diagnosis-related visual recognition tasks is studied. This study further validates the superiority of lesion synthesis ability through the proposed sample-adaptive learning framework over the common whole sample-space learning.

- In Chapter 6, the problem of significant visual variations among MR images is

identified for brain tumor segmentation tasks, which inevitably increases the difficulty of learning a well-performed unified CNN model. In the sample-adaptive learning based segmentation framework that is delicately established in this thesis, the different lesion contrasts varying with the input MR images are handled by the particularly learnt intensity adjustment. Utilizing the proposed sample-adaptive learning framework, every input image has its unique intensity transformation before processed by the subsequent segmentation network. Also, this transformation is learnt with the following segmentation network. This end-to-end learning strategy enhances the tumor-related contrast to best serve the segmentation task. The thesis demonstrates its effectiveness in promoting the brain tumor segmentation performance under the scenarios of both single and multiple modalities on two public datasets. Moreover, the generalization capacity of sample-adaptive learning in adjusting medical image intensities is further studied through transferring the specifically learnt transformation from one segmentation network to another. This study suggests that some general information about how to sample-adaptively mitigate the intensity variation among MR images for the segmentation task, rather than for a specific segmentation model, has been learnt.

7.2 Future Work

For the work in this thesis, its future research directions that could be further explored are discussed as follows:

- Semi-supervised learning. Semi-supervised learning has demonstrated its promising performance on various computer vision tasks. For medical images, as mentioned, scarce labeled data is always a problem, which hinders learning an effective prediction model. Semi-supervised learning can utilize only a small number of labeled data and also more unlabeled data to train the model. For example, in the cross-modality MR image synthesis task, since the scanning of two-modality images for every patient requires enough time and money, the collection is not easy. If a semi-supervised learning strategy could be explored in this case, the training of model will benefit from more visual information of various source-modality images so that the synthesis performance can be promoted.

- Weakly supervised learning. Weakly supervised learning aims to apply simpler/weaker ground-truth labels on a more complicated learning task. During the learning for medical image segmentation, the dense labels of training images should be provided in advance. However, getting the segmentation labels of medical images needs the manual annotation from the experts with professional medical backgrounds. This

inevitably increases the workload in clinics. Thus, only annotating the weaker labels rather than the dense voxel-wise ground-truths for training images is promising to mitigate this issue. There are two types of weak labels that may be suitable for the complex medical image segmentation task. The first type is the bounding box of target objects. Only providing the coarse locations of the interesting regions reduces much laborious effort. The second one is the scratches of target organs or lesions. Interactive annotations could also free the hands of experts. Therefore, applying weakly supervised learning by these simpler labels could guarantee the sufficient number of labeled data during training.

- Transfer learning. Transfer learning can exploit the deep models which are pre-trained in solving one computer vision problem to a different but related problem. It would be highly useful in the medical image segmentation cases. As aforementioned, the training of deep models for medical image segmentation always faces the situation of lacking sufficient labels, which has the adverse effects on model learning. Compared with the medical images, annotated generic images are generally more accessible. Thus, how to transfer the pre-trained generic image segmentation CNN models to medical image segmentation tasks is worthy to explore. Since it is a cross-domain and cross-task work, more efforts should be made to minimizing the distance between two-domain data and also extracting the relationship between two segmentation tasks. Besides, not only from the generic image segmentation but also from the other medical image segmentation tasks, transfer learning could utilize the deep models trained on a relatively sufficient medical image dataset to process the images from another dataset. For example, using the CNNs for the prevalent glioma segmentation task to initialize the deep models for the segmentation work of the other rarer lesions in brain.

- CT/PET datasets. In this thesis, the proposed medical image per-voxel prediction frameworks are validated on MR image datasets. Similar to MR images, CT/PET images have three spatial dimensions and image-wise visual characters especially when they are scanned on the lesion contained body parts. In addition to the proposed 3D cGAN in Chapter 3 already applied on PET datasets in our paper [129], our Ea-GANs, SA-GANs, and SA-LuT-Nets may be also effective for CT/PET per-voxel prediction tasks. Through the proposed Ea-GANs, the edge information in CT/PET images could be enhanced during the synthesis, so that the synthesized images will have sharper and clearer appearance. Via our SA-GANs and SA-LuT-Nets, the variation among CT/PET images could be considered. Thus, the uniquely learnt prediction model for each CT/PET sample may achieve better estimation performance.

Moreover, the first three future research directions can be integrated together to learn

the per-voxel prediction on medical images. The weakly supervised learning strategy could be cooperated with the semi-supervised learning approach, which will exploit a few densely labeled images and a large number of weakly annotated images to fully utilize the accessible visual information and ensure the segmentation performance. In addition, the semi-supervised learning and transfer learning could be incorporated into the model learning. Using both of the labeled images from other dataset and the scarce annotated medical images is highly potential to train a well-performed deep CNN model. The above research directions are expected to develop more effective and efficient medical image per-voxel prediction frameworks in the future.

Besides, the final goal of per-voxel prediction on medical images is to assist experts in disease diagnosis and treatment monitoring. However, most existing works in this field only focus on prediction, which means that the per-voxel prediction is regarded as an independent step from the final diagnosis/treatment. In the future, we could introduce the diagnosis/treatment information into the learning of per-voxel prediction. In this way, the task-specific prediction results can be more helpful for experts. Furthermore, an integration system, consisting of single-modality image scanning, multi-modality synthesis, object segmentation, disease diagnosis, and treatment monitoring, could be developed to adequately replace the manual work of experts. This automatic system may fundamentally solve the problem of sparse medical resources in today's world.

Bibliography

(1) Y. Bengio, A. Courville and P. Vincent, "Representation learning: A review and new perspectives", *IEEE transactions on pattern analysis and machine intelligence*, 2013, **35**, 1798–1828.

(2) A. Krizhevsky, I. Sutskever and G. E. Hinton, Advances in neural information processing systems, 2012, pp. 1097–1105.

(3) H. Noh, S. Hong and B. Han, Proceedings of the IEEE international conference on computer vision, 2015, pp. 1520–1528.

(4) O. M. Parkhi, A. Vedaldi and A. Zisserman, "Deep face recognition", 2015.

(5) G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken and C. I. Sánchez, "A survey on deep learning in medical image analysis", *Medical image analysis*, 2017, **42**, 60–88.

(6) M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin and H. Larochelle, "Brain tumor segmentation with deep neural networks", *Medical image analysis*, 2017, **35**, 18–31.

(7) J. Long, E. Shelhamer and T. Darrell, Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

(8) O. Ronneberger, P. Fischer and T. Brox, International Conference on Medical image computing and computer-assisted intervention, 2015, pp. 234–241.

(9) P. Isola, J.-Y. Zhu, T. Zhou and A. A. Efros, Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.

(10) X. Yi, E. Walia and P. Babyn, "Generative adversarial network in medical imaging: A review", *Medical image analysis*, 2019, 101552.

(11) M. Goetz, C. Weber, F. Binczyk, J. Polanska, R. Tarnawski, B. Bobek-Billewicz, U. Koethe, J. Kleesiek, B. Stieltjes and K. H. Maier-Hein, "DALSA: domain adaptation for supervised learning from sparsely annotated MR images", *IEEE transactions on medical imaging*, 2015, **35**, 184–196.

(12) M. Chen, K. Q. Weinberger and J. Blitzer, Advances in neural information processing systems, 2011, pp. 2456–2464.

(13) M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C. R. Guttmann, F.-E. de Leeuw, C. M. Tempany, B. van Ginneken et al., International conference on medical image computing and computer-assisted intervention, 2017, pp. 516–524.

(14) B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest et al., "The multimodal brain tumor image segmentation benchmark (BRATS)", *IEEE transactions on medical imaging*, 2015, **34**, 1993–2024.

(15) *The IXI dataset*, `http://brain-development.org/ixi-dataset/`, Accessed: 2020-05-06.

(16) O. Maier, B. H. Menze, J. von der Gablentz, L. Häni, M. P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen et al., "ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI", *Medical image analysis*, 2017, **35**, 250–269.

(17) F. Isensee, P. Kickingereder, W. Wick, M. Bendszus and K. H. Maier-Hein, International MICCAI Brainlesion Workshop, 2017, pp. 287–297.

(18) C. Chen, X. Liu, M. Ding, J. Zheng and J. Li, International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019, pp. 184–192.

(19) S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani and C. Davatzikos, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features", *Scientific data*, 2017, **4**, 170117.

(20) Q. Zhang, L. T. Yang, Z. Chen and P. Li, "A survey on deep learning for big data", *Information Fusion*, 2018, **42**, 146–157.

(21) Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, 1998, **86**, 2278–2324.

(22) V. Nair and G. E. Hinton, Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 807–814.

(23) C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

(24) K. He, X. Zhang, S. Ren and J. Sun, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

(25) G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

(26) L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs", *IEEE transactions on pattern analysis and machine intelligence*, 2017, **40**, 834–848.

(27) I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, Advances in neural information processing systems, 2014, pp. 2672–2680.

(28) I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin and A. C. Courville, Advances in neural information processing systems, 2017, pp. 5767–5777.

(29) M. Mirza and S. Osindero, "Conditional generative adversarial nets", *arXiv preprint arXiv:1411.1784*, 2014.

(30) P. Isola, J.-Y. Zhu, T. Zhou and A. A. Efros, "Image-to-image translation with conditional adversarial networks", *arXiv preprint arXiv:1611.07004*, 2016.

(31) J.-Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks", *arXiv preprint arXiv:1703.10593*, 2017.

(32) D. Sungatullina, E. Zakharov, D. Ulyanov and V. Lempitsky, Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 579–595.

(33) M. Hofmann, F. Steinke, V. Scheel, G. Charpiat, J. Farquhar, P. Aschoff, M. Brady, B. Scholkopf and B. J. Pichler, "MRI-based attenuation correction for PET/MRI: a novel approach combining pattern recognition and atlas registration", *Journal of nuclear medicine*, 2008, **49**, 1875.

(34) M. Hofmann, I. Bezrukov, F. Mantlik, P. Aschoff, F. Steinke, T. Beyer, B. J. Pichler and B. Schölkopf, "MRI-based attenuation correction for whole-body PET/MRI: quantitative evaluation of segmentation-and atlas-based methods", *Journal of Nuclear Medicine*, 2011, **52**, 1392–1399.

(35) S. Roy, A. Carass and J. L. Prince, "Magnetic resonance image example-based contrast synthesis", *IEEE transactions on medical imaging*, 2013, **32**, 2348–2363.

(36) N. Burgos, M. J. Cardoso, K. Thielemans, M. Modat, S. Pedemonte, J. Dickson, A. Barnes, R. Ahmed, C. J. Mahoney, J. M. Schott et al., "Attenuation correction synthesis for hybrid PET-MR scanners: application to brain studies", *IEEE transactions on medical imaging*, 2014, **33**, 2332–2341.

(37) M. Chen, A. Jog, A. Carass and J. L. Prince, Medical Imaging 2015: Image Processing, 2015, vol. 9413, 94131Q.

(38) T. Huynh, Y. Gao, J. Kang, L. Wang, P. Zhang, J. Lian and D. Shen, "Estimating CT image from MRI data using structured random forest and auto-context model", *IEEE transactions on medical imaging*, 2016, **35**, 174–183.

(39) Y. Wang, G. Ma, L. An, F. Shi, P. Zhang, D. S. Lalush, X. Wu, Y. Pu, J. Zhou and D. Shen, "Semisupervised Tripled Dictionary Learning for Standard-Dose PET Image Prediction Using Low-Dose PET and Multimodal MRI", *IEEE Transactions on Biomedical Engineering*, 2017, **64**, 569–579.

(40) D. H. Ye, D. Zikic, B. Glocker, A. Criminisi and E. Konukoglu, International Conference on Medical Image Computing and Computer-Assisted Intervention, 2013, pp. 606–613.

(41) A. Jog, A. Carass, S. Roy, D. L. Pham and J. L. Prince, "Random forest regression for magnetic resonance image synthesis", *Medical image analysis*, 2017, **35**, 475–488.

(42) H. Chen, Y. Zhang, W. Zhang, P. Liao, K. Li, J. Zhou and G. Wang, "Low-dose CT via convolutional neural network", *Biomedical optics express*, 2017, **8**, 679–694.

(43) H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou and G. Wang, "Low-dose CT with a residual encoder-decoder convolutional neural network", *IEEE transactions on medical imaging*, 2017, **36**, 2524–2535.

(44) E. Kang, J. Min and J. C. Ye, "A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction", *Medical physics*, 2017, **44**, e360–e375.

(45) K. Zeng, H. Zheng, C. Cai, Y. Yang, K. Zhang and Z. Chen, "Simultaneous single- and multi-contrast super-resolution for brain MRI images based on a convolutional neural network", *Computers in biology and medicine*, 2018, **99**, 133–141.

(46) A. S. Chaudhari, Z. Fang, F. Kogan, J. Wood, K. J. Stevens, E. K. Gibbons, J. H. Lee, G. E. Gold and B. A. Hargreaves, "Super-resolution musculoskeletal MRI using deep learning", *Magnetic resonance in medicine*, 2018, **80**, 2139–2154.

(47) D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang and D. Shen, "Medical Image Synthesis with Deep Convolutional Adversarial Networks", *IEEE Transactions on Biomedical Engineering*, 2018.

(48) L. Xiang, Y. Qiao, D. Nie, L. An, W. Lin, Q. Wang and D. Shen, "Deep auto-context convolutional neural networks for standard-dose PET image estimation from low-dose PET/MRI", *Neurocomputing*, 2017, **267**, 406–416.

(49) D. Nie, X. Cao, Y. Gao, L. Wang and D. Shen, in *Deep Learning and Data Labeling for Medical Applications*, Springer, 2016, pp. 170–178.

(50) X. Han, "MR-based synthetic CT generation using a deep convolutional neural network method", *Medical physics*, 2017, **44**, 1408–1419.

(51) A. P. Leynes, J. Yang, F. Wiesinger, S. S. Kaushik, D. D. Shanbhag, Y. Seo, T. A. Hope and P. E. Larson, "Direct pseudoCT generation for pelvis PET/MRI attenuation correction using deep convolutional neural networks with multi-parametric MRI: zero echo-time and dixon deep pseudoCT (ZeDD-CT)", *Journal of Nuclear Medicine*, 2017, jnumed–117.

(52) F. Liu, H. Jang, R. Kijowski, T. Bradshaw and A. B. McMillan, "Deep learning MR imaging–based attenuation correction for PET/MR imaging", *Radiology*, 2017, **286**, 676–684.

(53) A. Chartsias, T. Joyce, R. Dharmakumar and S. A. Tsaftaris, International Workshop on Simulation and Synthesis in Medical Imaging, 2017, pp. 3–13.

(54) H. Emami, M. Dong, S. P. Nejad-Davarani and C. K. Glide-Hurst, "Generating synthetic CTs from magnetic resonance images using generative adversarial networks", *Medical physics*, 2018, **45**, 3627–3636.

(55) Y. Hiasa, Y. Otake, M. Takao, T. Matsuoka, K. Takashima, A. Carass, J. L. Prince, N. Sugano and Y. Sato, International Workshop on Simulation and Synthesis in Medical Imaging, 2018, pp. 31–41.

(56) Z. Zhang, L. Yang and Y. Zheng, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9242–9251.

(57) A. Ben-Cohen, E. Klang, S. P. Raskin, M. M. Amitai and H. Greenspan, International Workshop on Simulation and Synthesis in Medical Imaging, 2017, pp. 49–57.

(58) L. Bi, J. Kim, A. Kumar, D. Feng and M. Fulham, in *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*, Springer, 2017, pp. 43–51.

(59) K. Armanious, C. Jiang, M. Fischer, T. Küstner, K. Nikolaou, S. Gatidis and B. Yang, "MedGAN: Medical image translation using GANs", *arXiv preprint arXiv:1806.06397*, 2018.

(60) H. Choi and D. S. Lee, "Generation of structural MR images from amyloid PET: Application to MR-less quantification", *Journal of Nuclear Medicine*, 2018, **59**, 1111–1117.

(61) W. Wei, E. Poirion, B. Bodini, S. Durrleman, N. Ayache, B. Stankoff and O. Colliot, International Conference on Medical Image Computing and Computer-Assisted Intervention, 2018, pp. 514–522.

(62) H. Van Nguyen, K. Zhou and R. Vemulapalli, International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015, pp. 677–684.

(63) A. Chartsias, T. Joyce, M. V. Giuffrida and S. A. Tsaftaris, "Multimodal mr synthesis via modality-invariant latent representation", *IEEE transactions on medical imaging*, 2018, **37**, 803–814.

(64) S. U. H. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem and T. Çukur, "Image Synthesis in Multi-Contrast MRI with Conditional Generative Adversarial Networks", *arXiv preprint arXiv:1802.01221*, 2018.

(65) S. Olut, Y. H. Sahin, U. Demir and G. Unal, International Workshop on PRedictive Intelligence In MEdicine, 2018, pp. 147–154.

(66) Q. Yang, N. Li, Z. Zhao, X. Fan, E. I. Chang, Y. Xu et al., "MRI Cross-Modality NeuroImage-to-NeuroImage Translation", *arXiv preprint arXiv:1801.06940*, 2018.

(67) P. Welander, S. Karlsson and A. Eklund, "Generative adversarial networks for image-to-image translation on multi-contrast MR images-A comparison of CycleGAN and UNIT", *arXiv preprint arXiv:1806.07777*, 2018.

(68) *The NAMIC dataset*, `https://http://hdl.handle.net/1926/1687`, Accessed: 2018-06-10.

(69) *The NBIA dataset*, `https://dcm.bmia.nl/ncia/login.jsf`, Accessed: 2020-05-06.

(70) C. Peterfy, E. Schneider and M. Nevitt, "The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee", *Osteoarthritis and cartilage*, 2008, **16**, 1433–1441.

(71) *The Brainweb dataset*, `https://brainweb.bic.mni.mcgill.ca/brainweb/`, Accessed: 2020-05-06.

(72) T. Zhou, H. Fu, G. Chen, J. Shen and L. Shao, "Hi-net: hybrid-fusion network for multi-modal MR image synthesis", *IEEE transactions on medical imaging*, 2020.

(73) *The low-dose CT dataset*, `https://www.aapm.org/GrandChallenge/LowDoseCT/`, Accessed: 2020-05-06.

(74) *The ADNI dataset*, `http://adni.loni.usc.edu/`, Accessed: 2020-05-06.

(75) E. Bullitt, D. Zeng, G. Gerig, S. Aylward, S. Joshi, J. K. Smith, W. Lin and M. G. Ewend, "Vessel tortuosity and brain tumor malignancy: a blinded study1", *Academic radiology*, 2005, **12**, 1232–1240.

(76) T. C. Mok and A. C. Chung, International MICCAI Brainlesion Workshop, 2018, pp. 70–80.

(77) *The MM-WHS dataset*, `http://www.sdspeople.fudan.edu.cn/zhuangxiahai/0/mmwhs/`, Accessed: 2020-05-06.

(78) D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium et al., "The WU-Minn human connectome project: an overview", *Neuroimage*, 2013, **80**, 62–79.

(79) A. Farag, L. Lu, H. R. Roth, J. Liu, E. Turkbey and R. M. Summers, "A bottom-up approach for pancreas segmentation using cascaded superpixels and (deep) image patch labeling", *IEEE Transactions on Image Processing*, 2016, **26**, 386–399.

(80) W. Li et al., "Automatic segmentation of liver tumor in CT images with deep convolutional neural networks", *Journal of Computer and Communications*, 2015, **3**, 146.

(81) T. Heimann, B. Van Ginneken, M. A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes et al., "Comparison and evaluation of methods for liver segmentation from CT datasets", *IEEE transactions on medical imaging*, 2009, **28**, 1251–1265.

(82) W. Thong, S. Kadoury, N. Piché and C. J. Pal, "Convolutional networks for kidney segmentation in contrast-enhanced CT scans", *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 2018, **6**, 277–282.

(83) P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. De Vries, M. J. Benders and I. Išgum, "Automatic segmentation of MR brain images with a convolutional neural network", *IEEE transactions on medical imaging*, 2016, **35**, 1252–1261.

(84) P. Moeskops, J. M. Wolterink, B. H. van der Velden, K. G. Gilhuijs, T. Leiner, M. A. Viergever and I. Išgum, International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016, pp. 478–486.

(85) H. Chen, Q. Dou, L. Yu, J. Qin and P.-A. Heng, "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images", *NeuroImage*, 2018, **170**, 446–455.

(86) A. M. Mendrik, K. L. Vincken, H. J. Kuijf, M. Breeuwer, W. H. Bouvy, J. De Bresser, A. Alansary, M. De Bruijne, A. Carass, A. El-Baz et al., "MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans", *Computational intelligence and neuroscience*, 2015, **2015**.

(87) H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey and R. M. Summers, International conference on medical image computing and computer-assisted intervention, 2015, pp. 556–564.

(88) M. Zreik, T. Leiner, B. D. De Vos, R. W. van Hamersvelt, M. A. Viergever and I. Išgum, 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), 2016, pp. 40–43.

(89) A. de Brebisson and G. Montana, Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2015, pp. 20–28.

(90) B. Landman and S. Warfield, Medical image computing and computer assisted intervention conference, 2012.

(91) P. V. Tran, "A fully convolutional neural network for cardiac segmentation in short-axis MRI", *arXiv preprint arXiv:1604.00494*, 2016.

(92) P. Radau, Y. Lu, K. Connelly, G. Paul, A. Dick and G. Wright, "Evaluation framework for algorithms segmenting short axis cardiac MRI", *The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge*, 2009, **49**.

(93) Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman and A. L. Yuille, International conference on medical image computing and computer-assisted intervention, 2017, pp. 693–701.

(94) Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin and P.-A. Heng, International conference on medical image computing and computer-assisted intervention, 2016, pp. 149–157.

(95) L. Yu, X. Yang, H. Chen, J. Qin and P. A. Heng, Thirty-first AAAI conference on artificial intelligence, 2017.

(96) G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang et al., "Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge", *Medical image analysis*, 2014, **18**, 359–373.

(97) R. P. Poudel, P. Lamata and G. Montana, in *Reconstruction, segmentation, and analysis of medical images*, Springer, 2016, pp. 83–94.

(98) Q. Zhu, B. Du, B. Turkbey, P. L. Choyke and P. Yan, 2017 International Joint Conference on Neural Networks (Ijcnn), 2017, pp. 178–184.

(99) F. Milletari, N. Navab and S.-A. Ahmadi, 3D Vision (3DV), 2016 Fourth International Conference on, 2016, pp. 565–571.

(100) J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers and I. B. Ayed, "HyperDenseNet: a hyper-densely connected CNN for multi-modal image segmentation", *IEEE transactions on medical imaging*, 2018, **38**, 1116–1126.

(101) L. Wang, D. Nie, G. Li, É. Puybareau, J. Dolz, Q. Zhang, F. Wang, J. Xia, Z. Wu, J.-W. Chen et al., "Benchmark on automatic six-month-old infant brain segmentation algorithms: the iSeg-2017 challenge", *IEEE transactions on medical imaging*, 2019, **38**, 2219–2230.

(102) H. Jia, Y. Xia, Y. Song, D. Zhang, H. Huang, Y. Zhang and W. Cai, "3D APA-Net: 3D adversarial pyramid anisotropic convolutional network for prostate segmentation in MR images", *IEEE transactions on medical imaging*, 2019.

(103) D. Yang, D. Xu, S. K. Zhou, B. Georgescu, M. Chen, S. Grbic, D. Metaxas and D. Comaniciu, International Conference on Medical Image Computing and Computer-Assisted Intervention, 2017, pp. 507–515.

(104) P. Moeskops, M. Veta, M. W. Lafarge, K. A. Eppenhof and J. P. Pluim, in *Deep learning in medical image analysis and multimodal learning for clinical decision support*, Springer, 2017, pp. 56–64.

(105) S. Valverde, M. Cabezas, E. Roura, S. González-Villà, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, A. Oliver and X. Lladó, "Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach", *NeuroImage*, 2017, **155**, 159–168.

(106) A. Carass, S. Roy, A. Jog, J. L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. H. Sudre et al., "Longitudinal multiple sclerosis lesion segmentation: resource and challenge", *NeuroImage*, 2017, **148**, 77–102.

(107) K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation", *Medical image analysis*, 2017, **36**, 61–78.

(108) S. Roy, J. A. Butman, D. S. Reich, P. A. Calabresi and D. L. Pham, "Multiple sclerosis lesion segmentation from brain MRI via fully convolutional neural networks", *arXiv preprint arXiv:1803.09172*, 2018.

(109) P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser et al., "The liver tumor segmentation benchmark (lits)", *arXiv preprint arXiv:1901.04056*, 2019.

(110) P. F. Christ, M. E. A. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, M. Armbruster, F. Hofmann, M. D'Anastasi et al., International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016, pp. 415–423.

(111) T. Nair, D. Precup, D. L. Arnold and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation", *Medical image analysis*, 2020, **59**, 101557.

(112) F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, International MICCAI Brainlesion Workshop, 2018, pp. 234–244.

(113) A. Myronenko, International MICCAI Brainlesion Workshop, 2018, pp. 311–320.

(114) W. Chen, B. Liu, S. Peng, J. Sun and X. Qiao, International MICCAI Brainlesion Workshop, 2018, pp. 358–368.

(115) K. Kamnitsas, W. Bai, E. Ferrante, S. McDonagh, M. Sinclair, N. Pawlowski, M. Rajchl, M. Lee, B. Kainz, D. Rueckert et al., International MICCAI Brainlesion Workshop, 2017, pp. 450–462.

(116) C. Zhou, S. Chen, C. Ding and D. Tao, International MICCAI Brainlesion Workshop, 2018, pp. 497–507.

(117) Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity", *IEEE transactions on image processing*, 2004, **13**, 600–612.

(118) L. Clarke, R. Velthuizen, M. Camacho, J. Heine, M. Vaidyanathan, L. Hall, R. Thatcher and M. Silbiger, "MRI segmentation: methods and applications", *Magnetic resonance imaging*, 1995, **13**, 343–368.

(119) C. R. Jack, R. C. Petersen, Y. C. Xu, P. C. O'Brien, G. E. Smith, R. J. Ivnik, B. F. Boeve, S. C. Waring, E. G. Tangalos and E. Kokmen, "Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment", *Neurology*, 1999, **52**, 1397–1397.

(120) M. Dadar, T. A. Pascoal, S. Manitsirikul, K. Misquitta, V. S. Fonov, M. C. Tartaglia, J. Breitner, P. Rosa-Neto, O. T. Carmichael, C. Decarli et al., "Validation of a regression technique for segmentation of white matter hyperintensities in Alzheimer's disease", *IEEE transactions on medical imaging*, 2017, **99**, 1–1.

(121) M. Lê, H. Delingette, J. Kalpathy-Cramer, E. R. Gerstner, T. Batchelor, J. Unkelbach and N. Ayache, "Personalized radiotherapy planning based on a computational tumor growth model", *IEEE transactions on medical imaging*, 2017, **36**, 815–825.

(122) X. Yi and P. Babyn, "Sharpness-aware Low dose CT denoising using conditional generative adversarial network", *arXiv preprint arXiv:1708.06453*, 2017.

(123) D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang and D. Shen, International Conference on Medical Image Computing and Computer-Assisted Intervention, 2017, pp. 417–425.

(124) Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox and O. Ronneberger, International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016, pp. 424–432.

(125) P. Costa, A. Galdran, M. I. Meyer, M. Niemeijer, M. Abràmoff, A. M. Mendonça and A. Campilho, "End-to-end adversarial retinal image synthesis", *IEEE transactions on medical imaging*, 2018, **37**, 781–791.

(126) H. Zhao, H. Li, S. Maurer-Stroh and L. Cheng, "Synthesizing retinal and neuronal images with generative adversarial nets", *Medical image analysis*, 2018, **49**, 14–26.

(127) J. T. Guibas, T. S. Virdi and P. S. Li, "Synthetic Medical Images from Dual Generative Adversarial Networks", *arXiv preprint arXiv:1709.01872*, 2017.

(128) J. M. Wolterink, A. M. Dinkla, M. H. Savenije, P. R. Seevinck, C. A. van den Berg and I. Išgum, "MR-to-CT Synthesis using Cycle-Consistent Generative Adversarial Networks", 2017.

(129) Y. Wang, B. Yu, L. Wang, C. Zu, D. S. Lalush, W. Lin, X. Wu, J. Zhou, D. Shen and L. Zhou, "3D conditional generative adversarial networks for high-quality PET image estimation at low dose", *NeuroImage*, 2018, **174**, 550–562.

(130) B. Yu, L. Zhou, L. Wang, J. Fripp and P. Bourgeat, Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on, 2018, pp. 626–630.

(131) F. Calimeri, A. Marzullo, C. Stamile and G. Terracina, International Conference on Artificial Neural Networks, 2017, pp. 626–634.

(132) C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda, S. Muramatsu, Y. Furukawa, G. Mauri and H. Nakayama, Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on, 2018, pp. 734–738.

(133) Y. Huo, Z. Xu, S. Bao, A. Assad, R. G. Abramson and B. A. Landman, Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on, 2018, pp. 1217–1220.

(134) M. Mardani, E. Gong, J. Y. Cheng, S. Vasanawala, G. Zaharchuk, M. Alley, N. Thakur, S. Han, W. Dally, J. M. Pauly et al., "Deep generative adversarial networks for compressed sensing automates MRI", *arXiv preprint arXiv:1706.00051*, 2017.

(135) Y. Hu, E. Gibson, L.-L. Lee, W. Xie, D. C. Barratt, T. Vercauteren and J. A. Noble, in *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*, Springer, 2017, pp. 105–115.

(136) F. Mahmood, R. Chen and N. J. Durr, "Unsupervised reverse domain adaption for synthetic medical images via adversarial training", *arXiv preprint arXiv:1711.06606*, 2017.

(137) K. Suzuki, I. Horiba and N. Sugie, "Neural edge enhancer for supervised edge enhancement from noisy images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003, **25**, 1582–1596.

(138) H. Khotanlou, O. Colliot, J. Atif and I. Bloch, "3D brain tumor segmentation in MRI using fuzzy classification, symmetry analysis and spatially constrained deformable models", *Fuzzy sets and systems*, 2009, **160**, 1457–1473.

(139) A. Bustin, D. Voilliot, A. Menini, J. Felblinger, C. de Chillou, D. Burschka, L. Bonnemains and F. Odille, "Isotropic Reconstruction of MR Images using 3D Patch-Based Self-Similarity Learning", *IEEE Transactions on Medical Imaging*, 2018.

(140) T. Sugahara, Y. Korogi, M. Kochi, I. Ikushima, Y. Shigematu, T. Hirai, T. Okuda, L. Liang, Y. Ge, Y. Komohara et al., "Usefulness of diffusion-weighted MRI with echo-planar technique in the evaluation of cellularity in gliomas", *Journal of magnetic resonance imaging*, 1999, **9**, 53–60.

(141) T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford and X. Chen, Advances in Neural Information Processing Systems, 2016, pp. 2234–2242.

(142) I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks", *arXiv preprint arXiv:1701.00160*, 2016.

(143) Y. Huang, L. Shao and A. F. Frangi, "Cross-Modality Image Synthesis via Weakly Coupled and Geometry Co-Regularized Joint Dictionary Learning", *IEEE transactions on medical imaging*, 2018, **37**, 815–827.

(144) K.-H. Thung, P.-T. Yap, E. Adeli, S.-W. Lee and D. Shen, "Conversion and time-to-conversion predictions of mild cognitive impairment using low-rank affinity pursuit denoising and matrix completion", *Medical image analysis*, 2018, **45**, 68–82.

(145) L. Fang, L. Zhang, D. Nie, X. Cao, I. Rekik, S.-W. Lee, H. He and D. Shen, "Automatic Brain Labeling via Multi-Atlas Guided Fully Convolutional Networks", *Medical image analysis*, 2018.

(146) R. Tyleček and R. Šára, German Conference on Pattern Recognition, 2013, pp. 364–374.

(147) M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth and B. Schiele, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3213–3223.

(148) B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp and P. Bourgeat, "Ea-GANs: Edge-aware Generative Adversarial Networks for Cross-modality MR Image Synthesis", *IEEE transactions on medical imaging*, 2019.

(149) C. Liu, J. Yuen and A. Torralba, "Nonparametric scene parsing via label transfer", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, **33**, 2368–2382.

(150) A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless and D. H. Salesin, Proceedings of the 28th annual conference on Computer graphics and interactive techniques, 2001, pp. 327–340.

(151) X. Wang, R. Girshick, A. Gupta and K. He, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

(152) A. Işın, C. Direkoğlu and M. Şah, "Review of MRI-based brain tumor image segmentation using deep learning methods", *Procedia Computer Science*, 2016, **102**, 317–324.

(153) S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge", *arXiv preprint arXiv:1811.02629*, 2018.

(154) L. G. Nyúl and J. K. Udupa, "On standardizing the MR image intensity scale", *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 1999, **42**, 1072–1081.

(155) L. G. Nyúl, J. K. Udupa and X. Zhang, "New variants of a method of MRI scale standardization", *IEEE transactions on medical imaging*, 2000, **19**, 143–150.

(156) S. Pereira, A. Pinto, V. Alves and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images", *IEEE transactions on medical imaging*, 2016, **35**, 1240–1251.

(157) M. Shah, Y. Xiao, N. Subbanna, S. Francis, D. L. Arnold, D. L. Collins and T. Arbel, "Evaluating intensity normalization on MRIs of human brain with multiple sclerosis", *Medical image analysis*, 2011, **15**, 267–282.

(158) D. C. Castro and B. Glocker, International Conference on Medical Image Computing and Computer-Assisted Intervention, 2018, pp. 206–214.

(159) M. Ju, C. Ding, Y. J. Guo and D. Zhang, "Idgcp: Image dehazing based on gamma correction prior", *IEEE Transactions on Image Processing*, 2019, **29**, 3104–3118.

(160) H. Farid, "Blind inverse gamma correction", *IEEE transactions on image processing*, 2001, **10**, 1428–1433.

(161) S.-C. Huang, F.-C. Cheng and Y.-S. Chiu, "Efficient contrast enhancement using adaptive gamma correction with weighting distribution", *IEEE transactions on image processing*, 2012, **22**, 1032–1041.

(162) F. Kallel and A. B. Hamida, "A new adaptive gamma correction based algorithm using DWT-SVD for non-contrast CT image enhancement", *IEEE transactions on nanobioscience*, 2017, **16**, 666–675.

(163) Y. Liang, L. Yang and H. Fan, 2009 International Conference on Optical Instruments and Technology: Optoelectronic Imaging and Process Technology, 2009, vol. 7513, 75130K.

(164) H. Xu, H. Xie, Y. Liu, C. Cheng, C. Niu and Y. Zhang, International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019, pp. 420–428.

(165) Y. Wang, Y. Zhou, W. Shen, S. Park, E. K. Fishman and A. L. Yuille, "Abdominal multi-organ segmentation with organ-attention networks and statistical fusion", *Medical image analysis*, 2019, **55**, 88–102.

(166) Q. Jin, Z. Meng, C. Sun, L. Wei and R. Su, "RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans", *arXiv preprint arXiv: 1811.01328*, 2018.

(167) P.-Y. Kao, T. Ngo, A. Zhang, J. W. Chen and B. Manjunath, International MICCAI Brainlesion Workshop, 2018, pp. 128–141.

(168) F. Wang, R. Jiang, L. Zheng, B. Biswal and C. Meng, "Brain-wise Tumor Segmentation and Patient Overall Survival Prediction", *arXiv preprint arXiv:1909.12901*, 2019.

(169) X. Li, G. Luo and K. Wang, "Multi-step Cascaded Networks for Brain Tumor Segmentation", *arXiv preprint arXiv:1908.05887*, 2019.

(170) Y. Xue, M. Xie, F. Farhat, O. Boukrina, A. Barrett, J. R. Binder, U. W. Roshan and W. W. Graves, "A multi-path decoder network for brain tumor segmentation", *submitted to BraTS (2019)*, 2019.

(171) Y.-X. Zhao, Y.-M. Zhang, M. Song and C.-L. Liu, International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019, pp. 256–265.